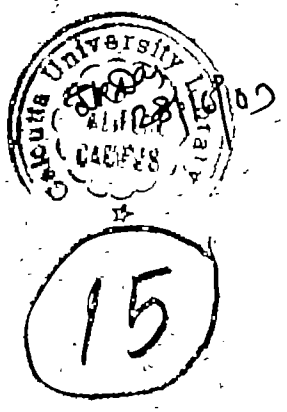
THE JOURNAL OF PHILOSOPHY

CUK - H06000 1 - 15 - KP 225



VOLUME C, NUMBER 7
JULY 2003

page

329 *Compliance, Complicity, and the Nature of Nonideal Conditions* Tamar Schapiro

356 *Some Problems for Conditionalization and Reflection* Frank Arntzenius

371 *The Stalnaker-Lewis Approach to Counterfactuals* Michael Tooley

378 NEW BOOKS: ANTHOLOGIES

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

You may also visit our website at: www.journalofphilosophy.org

100
J826

THE JOURNAL OF PHILOSOPHY

2003

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset and printed by Capital City Press, Montpelier, VT.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

27 5th
N 9

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U S Copyright Law. This consent does not extend to any other kinds of copying More information on our terms may be obtained by consulting our January issue or by writing to us

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices

POSTMASTER: Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2003 by the Journal of Philosophy, Inc.

ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY

VOLUME C, NO. 7, JULY 2003

COMPLIANCE, COMPLICITY, AND THE NATURE OF NONIDEAL CONDITIONS*

The question I address in this article is whether wrongdoing on the part of others can have the effect of mitigating the stringency with which moral principles apply to us. Can the fact that others behave badly make it the case that we are permitted to act in ways that would be impermissible but for such wrongdoing? I believe most of us are of two minds about this question. On the one hand, we think that it is never permissible to violate our moral standards in response to others' misconduct, because to do so is simply to "stoop to their level." Two wrongs, it would seem, just cannot make a right. On the other hand, we recognize that insisting on strict adherence to high-minded standards of conduct in the face of others' wrongdoing can be naïve at best, and at worst tragically misguided. To this extent we believe that it can be necessary, and indeed morally necessary, to fight fire with fire.

Kant is generally thought of as a purist, one who adheres unflinchingly to the first of these intuitions. There is evidence of this, of course, in his notorious response to Benjamin Constant, where Kant[1]

*I have presented this article at the University of California/Berkeley, Stanford University, the University of San Diego Values Institute (Conference on Kantian Ethics), the University of California/Davis, and the University of California/Santa Cruz. For comments on recent versions, I am indebted to my colleagues at Stanford, especially Allen Wood, Debra Satz, Michael Bratman, and Manuel Vargas. I am also grateful to Anthony Laden, Seana Shiffrin, Jay Wallace, Christopher Kutz, Aaron James, Elijah Millgram, Sarah Buss, Connie Rosati, Tom Hill, Barbara Herman, Talbot Brewer, John Doris, and James Mahon, whose responses I have found very instructive. For comments on earlier drafts, I thank Arthur Kuflik, Tamar Gendler, Michael Stocker, and several anonymous reviewers. I am also indebted to Dmitri Petrov, who provided valuable criticism at all points along the way.

[1] "On a Supposed Right to Lie from Philanthropy" (1797), included in the *Practical Philosophy* volume of *The Cambridge Edition of the Works of Immanuel Kant*, Paul Guyer and Allen Wood, eds. (New York: Cambridge, 1996), pp. 611–15/8:425–30. Unless otherwise noted, all citations of Kant will refer to this volume, with page references to this volume listed first, followed by references to the standard Academy Edition.

0022-362X/03/0007/329–55 © 2003 The Journal of Philosophy, Inc.

maintains that it is impermissible to lie to a murderer even for the sake of keeping him from his intended victim. Moreover, Kant's biting of the bullet in this particular case seems to be motivated by a fundamental feature of his theory, namely his conception of the moral law as a categorically binding imperative. Consider the following passage from the *Groundwork*:

> ...every rational being must act as if he were by his maxims at all times a lawgiving member of the universal kingdom of ends.... Now, such a kingdom of ends would actually come into existence through maxims whose rule the categorical imperative prescribes to all rational beings *if they were universally followed*. It is true that, even though a rational being scrupulously follows this maxim himself, he cannot for that reason count upon every other to be faithful to the same maxim.... [N]evertheless that law...remains in full force because it commands categorically. And just in this lies the paradox that the mere dignity of humanity as rational nature, without any other end or advantage to be attained by it—hence respect for a mere idea—is yet to serve as an inflexible precept of the will....[2]

To think that others' wrongdoing can in any way mitigate the stringency with which the moral law applies to us is, presumably, to deny its categorical status. For others' wrongdoing can only affect us by altering the causal nexus within which we act. That nexus determines the consequences of our own lawful action. But to think that the value of our own lawful action depends upon its consequences is simply to think of the law as a hypothetical imperative. Given that the moral law is categorically rather than hypothetically binding, others' misconduct cannot undermine its normative force.

There are, however, conflicting strains in Kant's corpus. The following passage from his *Lectures on Ethics*, for example, contains an expression of a more pragmatic position:

> If we were to be at all times punctiliously truthful we might often become victims of the wickedness of others who were ready to abuse our truthfulness. If all men were well-intentioned it would not only be a duty not to lie, but no one would do so because there would be no point in it. But as men are malicious, it cannot be denied that to be punctiliously truthful is often dangerous...if I cannot save myself by maintaining silence, then my lie is a weapon of defense.[3]

[2] *Groundwork of the Metaphysics of Morals* (1785), pp. 87–88/4:438–39

[3] *Lectures on Ethics*, Ak. 228, as quoted in Christine Korsgaard, "The Right to Lie. Kant on Dealing with Evil," in her collection *Creating the Kingdom of Ends* (New York: Cambridge, 1996), p. 144.

The image of a lie as a weapon of defense is certainly intelligible, but it is in tension with the central Kantian claim that the value of right action—and presumably the disvalue of wrong action—lies in the form of its maxim rather than its fitness to produce a certain result. How, then, can the more pragmatic position expressed in this passage be reconciled with Kant's more familiar purism?

I. THE PROBLEM

Rather than pursue this as an exegetical question particular to Kant, I want to ask more generally what structure a nonconsequentialist moral theory has to have if it is to make sense of both purist and pragmatic intuitions with regard to the wrongdoing of others. Christine Korsgaard has attempted to do the same, most notably in her article "The Right to Lie: Kant on Dealing with Evil" (*ibid.*). There Korsgaard argues that Kant could have offered a limited justification of the more pragmatic position by appealing to a distinction between "ideal" and "nonideal" conditions. On this account, nonideal conditions are defined as conditions under which the attempt to abide by the moral law in its most stringent form "would make you a tool of evil," and this consideration is supposed to justify your acting on a more permissive version of that law under such circumstances (*ibid.*, p. 153). But exactly how is the justification supposed to work? If the sense in which you are "made a tool of evil" is merely that you are made to bring about a very bad result as a consequence of your truthtelling, then it would seem the moral value of that action depends not on its lawful form but on the amount of good or evil it brings about. Although Korsgaard would clearly reject this interpretation of her position, her own account of it is sufficiently unclear to leave the worry open.[4] And without an alternative interpretation of what it is

[4] On the one hand, Korsgaard defines nonideal conditions as those under which the attempt to conform to the moral law in its most stringent form would make you a tool of evil (p. 153). On the other hand, she seems to agree that Kant's view seems to make the very existence of this sort of condition an impossibility. This is because, as Korsgaard explains, Kant holds a "clean hands" view of moral responsibility, according to which, as long as you do what is required of you by the moral law, you cannot be held accountable for any bad consequences which might result from your lawful action (p. 148). But this suggests that your compliance with the moral law can never in principle count as complicity with evil. If nonideal conditions are defined as those in which your compliance would count as complicity, then the clean hands view of responsibility rules out the possibility of Kantian nonideal theory. If the definition of nonideal conditions is supposed to provide a reason for mitigating the clean hands view, however, (and Korsgaard, on p. 150, suggests this is its role) then we need a different account of what the nonideality of nonideal conditions consists in. On p. 148, Korsgaard offers a gloss on Rawls's definition of nonideal conditions as conditions which "exist when, or to the extent that, the special conception of justice cannot be realized effectively." Characterizing the defect in this way, however, makes it sound as if the essential problem is one of inefficacy; nonideal conditions

to be made a tool of evil, the appeal to nonideal conditions can appear to be an ad hoc way of limiting the scope of Kant's purism.

That said, I do believe the general strategy of limiting purism by appeal to some conception of nonideal conditions is promising. What remains to be seen is whether it is possible to develop a conception of what the nonideality of nonideal conditions could consist in, such that those conditions could play the mitigating role claimed for them within a thoroughly nonconsequentialist moral theory.[5] My primary

---

are those under which adherence to ideal principles is somehow instrumentally defective. But the existence of this sort of defect could not provide grounds for departing from ideal principles on a truly nonconsequentialist account of such principles. Hence my dissatisfaction with the account of the nonideality of nonideal conditions as it appears in the "Right to Lie" article. That said, I believe Korsgaard implicitly revises this account, though without explicit mention of the ideal/nonideal distinction, in her later paper, "Taking the Law into Our Own Hands: Kant on the Right to Revolution," in *Reclaiming the History of Ethics: Essays for John Rawls*, Andrews Reath, Barbara Herman, and Korsgaard, eds. (New York: Cambridge, 1997), pp. 297–328, to which I am greatly indebted. I believe most of what I have to say in this paper is consistent with Korsgaard's position in this later paper, though that paper is still, I think, too vague about what the criterion is for determining whether or not a conscientious agent has reason to "take the law into his own hands" (cf. note 28).

[5] Readers familiar with Liam Murphy's work will recognize that I am setting up the problem in a different way than he does. Murphy starts out by defining nonideal conditions as conditions under which there is some significant amount of noncompliance, and then asking whether this condition has any bearing on the extent of our responsibilities. The answer to that question may turn out to be yes or no, and indeed Murphy's answer is no—the responsibilities we have under conditions of partial compliance are not significantly different from the responsibilities we have under conditions of full compliance—see his *Moral Demands in Nonideal Theory* (New York: Oxford, 2000). By contrast, I am assuming that the concept of a nonideal condition is the concept of a mitigating condition, a condition the existence of which has the effect of undermining the stringency with which some set of moral or political principles applies to us. I am asking what has to be true of such conditions for them to play this role, and I am using the condition of partial compliance as a test case to try to answer this question because this is one of the conditions which John Rawls identifies as "nonideal" in his original use of the concept—see Rawls, *A Theory of Justice* (Cambridge: Harvard, 1971), especially sections 2 and 39. There are two main reasons why I think the notion of nonideality should not simply be identified with the notion of partial compliance. One is that partial compliance is only one type of nonideal condition, according to Rawls, the other being the existence of "natural limitations and historical contingencies"—*A Theory of Justice*, p. 246. So, the nonideality of nonideal conditions cannot consist in the existence of noncompliance. Second, Rawls's criterion for drawing the line between the ideal condition of full compliance and the nonideal condition of partial compliance is not simply that under full compliance, no one violates the principles of justice, whereas under partial compliance, some people do. Indeed Rawls states that under full compliance, the violations which do occur "are counted as exceptions"—*A Theory of Justice*, p. 245. Thus, there must be some principle according to which we judge whether or not the degree or character of noncompliance is significant enough to not be discountable in this way. I take it that in order to find this principle, we need a prior account of what the nonideality of nonideal conditions consists in. I am grateful to Seana Shiffrin for urging me to be clearer on this point.

aim here is to lay out *the beginnings* of such an account, by examining a model of a type of rule-governed activity that would allow for an ideal/nonideal distinction of the right kind. That activity is a social practice, an activity governed by constitutive rules that apply to individuals in virtue of their occupancy of certain roles or positions. Within a practice, I will argue, the wrongdoing of others can pose a threat to us not simply by making it likely that we will, through our lawful activity, produce a bad outcome, but rather by undermining the integrity of that lawful activity. This is so, I will claim, because within a practice, our actions depend *constitutively* on others' compliance with the practice, so that beyond a certain threshold, others' wrongdoing can alter the character of what we ourselves are doing, depriving it of its status as an instance of participation in the form of activity defined by the practice rules. If this is right, then it is in this sense that nonideal conditions, at least with respect to a set of practice rules, can be defined as those under which the wrongdoing of others threatens to make us tools of evil. The further question, which I will not be able to address here, is whether action on the moral law can be construed analogously.

One additional note: apart from its implications for Kant's moral theory, I believe the account of nonideal conditions I offer here in connection with practice rules sheds light on the possibility and structure of a certain kind of moral dilemma that Thomas Nagel has called a "blind alley."[6] A blind alley is a situation in which an ideally virtuous agent is forced, through no fault of his own, to act in a way that is insufficiently justified. Nagel believes that such a predicament is possible, because he believes we are at bottom torn by two inescapable yet irreconcilable moral standards—one deontological and the other utilitarian. On Nagel's view, many instances of the tension

[6] "War and Massacre," in his collection, *Mortal Questions* (New York: Cambridge, 1979), pp. 53–74, here p. 74. This essay, along with Michael Walzer's "Political Action: The Problem of Dirty Hands," *Philosophy and Public Affairs*, II (1973): 160–80, gave rise to a significant body of literature on the question whether "moral dilemmas" are possible. For a basic anthology on the topic, see Christopher Gowans, ed., *Moral Dilemmas* (New York: Oxford, 1987). My claim here is to give an account of the origin of a certain kind of moral dilemma, though there may be others. One type of moral conflict I do not purport to shed light on is the kind that results from an overload of obligations. A person might experience such an overload because she occupies two distinct roles (say, that of parent and that of philosopher), or because she occupies one role to which a multitude of obligations attach (for example, University Provost). I leave it open how best to understand these types of conflicts, and whether or not they have the features of a "blind alley." As will become clear in what follows, the type of conflict I describe results not from overload, but from corruption of the background conditions necessary for obligations to sustain their binding force.

between purism and pragmatism can be accounted for in terms of this fundamental conflict. Now part of the appeal of Nagel's account, I take it, is that it seems to fully appreciate the depth and tenacity of certain types of moral conflict. This is in contrast to what Bernard Williams has called the "glib moralism" of theories which contend that the sense of moral conflict is always attributable to some error or shortcoming on the part of the agent.[7] I will claim, however, that an agent committed exclusively to a nonconsequentialist ideal of conduct can, under nonideal conditions, be faced with a predicament that has all the relevant characteristics of a blind alley. If this is so, then it is possible to avoid glib moralism without giving up on the unity of moral theory.

II. ACTION AS INTERACTION: THE CASE OF SOCIAL PRACTICES

II.1. Rawls on practice rules. The conception of a social practice which I will use as my starting point comes from John Rawls's early article, "Two Concepts of Rules."[8] There Rawls draws on a distinction, originally put forth by H.L.A. Hart, between nonconstitutive and constitutive rules.[9] In Rawls's terminology, the distinction is between "summary" rules, which we normally call "rules of thumb," and "practice rules," which define games, institutions, and the like. Summary rules are generalized reports of the results of applying some other rule directly to the cases at hand. For example, the general rules employed at the middle level of a utilitarian theory are regarded as summary rules when they are thought of as approximating the results of applying the principle of utility case by case. But the use of summary rules is not particular to utilitarianism. Indeed, Kant's duties to self and others can also be regarded as summary rules, generalizations of the results of applying the moral law to particular maxims, case by case. Practice rules, by contrast, are not arrived at by generalization, and do not function as reports of the results of some prior procedure. Rather, they define procedures compliance with which constitutes participation in some new form of activity. By complying with the rules of baseball, for example, a participant in that practice makes his behavior count as a move in that game (for example, hitting a

[7] "Ethical Consistency," in *Problems of the Self: Philosophical Papers 1956–1972* (New York: Cambridge, 1973), pp. 166–86, here p. 173.

[8] Originally published in *Philosophical Review*, LXIV, 1 (January 1955). 3–32; reprinted in Rawls, *Collected Papers*, Samuel Freeman, ed. (Cambridge: Harvard, 1999), pp. 20–46. Citations will be from the latter volume.

[9] *The Concept of Law* (New York: Oxford, 1961). See his distinction between "laws that confer powers" and those that "impose duties and are analogous to orders backed by threats," p. 32 of the paperback edition (New York: Oxford, 1997).

"home run"). Rawls's claim is that such a move is, *in the first instance*, a piece of behavior that is only properly described in terms of the practice rules that govern it. Outside the "stage setting" of the game of baseball, for example, it is indeed physically possible to hit a hard, leather-covered sphere with a strangely-shaped piece of wood so that it travels a certain distance along a certain trajectory, but it is not logically possible for that bit of behavior to count as "hitting a home run." The same is not the case with actions falling under summary rules, the proper descriptions of which have to be assumed to be fixed prior to the application of the summary rule. In this sense, Rawls maintains, practice rules are logically prior to the actions falling under them, whereas summary rules are not (*op. cit.*, p. 37).

Rawls's main point in "Two Concepts of Rules" is that the justification of an action falling under a practice is different from the justification of the practice itself (*op. cit.*, p. 31). Because the actions falling under practice rules are logically constituted by those rules, such actions can only be justified by being shown to be in accordance with those rules. To justify a move by showing its conformity to some standard that is independent of the game is to justify it *as* some other, practice-independent form of behavior, rather than as the move that it is. Rawls applies this insight to argue that participants in practices that are governed by standards other than the principle of utility cannot, without logical incoherence, justify their actions by appeal to the principle of utility. He claims, for example, that one who occupies the role of "promisor" cannot justify keeping a promise on the grounds that doing so would promote the most good overall, because to do so would simply be to reveal that the person does not understand what promising is. The only justification of promise-keeping that he can offer, Rawls maintains, is one that appeals to the constitutive rules of promising (*op. cit.*, p. 39). Moreover if he wants to justify breaking a promise, he can only do so by appealing to one of the various excusing conditions built into the content of the practice rules themselves (*op. cit.*, p. 32).

Although the justification of particular actions falling under practice rules cannot make appeal to practice-independent standards, Rawls maintains that the justification of a practice as a whole must appeal to such standards. For if there is any reason to set up a practice in the first place, Rawls claims, that reason must be based on a principle that is defined prior to the practice. Hence the standpoint of what I will call the *social engineer* is one that can be governed directly by the principle of utility, even though the standpoint of the participant

cannot (*op. cit.*, p. 31).[10] By the same token, a participant who wishes to make changes in the practice as a whole cannot do so without stepping out of his role as a participant and instead acting from the standpoint of a social engineer.[11] Hence one cannot justify an act of, say, promise-breaking on the grounds that the rules of promising are not serving their practice-independent purpose. That can be a justification for changing the rules, but it cannot be a justification for breaking them.

II.2. When a practice becomes a sham. Rawls's aim in making this logical point is to strengthen the case for a two-level version of utilitarianism. His claim is that if general rules are construed as practice rather than summary rules, utilitarianism need not permit counterintuitive violations of our ordinary standards, violations such as punishing the innocent for the sake of the greater good. Later on, I will have something to say about whether this defense of utilitarianism is successful. But for now I am simply going to draw on what I take to be Rawls's valid and illuminating analysis of practice rules in order to show—or rather to *begin* to show—how it can provide support for two-level Kantianism. As such, my strategy is to put Rawls's notion of practice rules to a use that is, in a sense, directly opposed to the use he originally made of it in "Two Concepts of Rules." Instead of arguing that the notion of a practice rule can be invoked to build the right kind of fixity into a two-level utilitarian theory, I will argue (or rather, provide the first step of an argument) that that notion can be invoked to build the right kind of flexibility into a two-level Kantian theory.

Rawls's argument focuses on examples where an agent fails to follow the rules of a given practice, choosing her action instead on the basis of some practice-independent standard, while still attempting to justify what she does by appeal to the practice rules. His point is that this results in a kind of constitutive and justificatory failure—having chosen the action on the basis of practice-independent principles, the action she in fact performs is not properly described as a move in the practice, and hence the practice rules cannot be invoked to justify *it*. The cases I will be focusing on have the same structure, except that the source of constitutive and justificatory failure is external to the agent's will. Suppose you are a conscientious participant in a practice, and you choose to perform a certain action on the basis of

[10] Rawls does not explicitly use the concept of a "social engineer." I am using this term so as to be able, later, to draw a distinction between the standpoint of the "social engineer" and that of the "reformer."

[11] This claim will require qualification later, when I discuss the role of the "reformer."

your judgment that doing so is in conformity with the practice rules.[12] Suppose further that your judgment is correct—the practice rules do demand this sort of conduct in this sort of situation. Moreover, nothing prevents you from acting on your decision, so you go ahead and do so. Have you thereby made the move?

While it may seem obvious that you have, this is not necessarily the case. Consider a simple example. You are a party to a dispute, and you are negotiating with the other party. Let us assume that negotiation is a practice, and that it is constitutive of this practice that actions falling under it be regulated by an ideal of "good faith"; that is, one who is not at least approximating this regulative ideal is not, strictly speaking, negotiating. Let us assume further that you are indeed negotiating in good faith: you put forth your claims as reasonably as you can, and you try to listen to the considerations of the other side as impartially as you can, in the interest of reaching a mutually acceptable agreement. During the course of the negotiation, however, it becomes clear to you that the representative of the other side is not governing himself the way you are. What he wants to do is to give the appearance that he is negotiating, while stalling for time. Outwardly, he goes through the motions of listening and responding to your arguments, but you are aware that he is really just trying to manage you, to work around you like an obstacle, without really engaging with the substance of your claims. Given this situation, what is the proper way to describe *your own* actions? Are you negotiating, or are you just babbling on?

If it makes sense to raise this question, this suggests that we take there to be some threshold beyond which the noncompliance of another can have a bearing on the integrity, and not merely the efficiency, of your rule-governed action. Let us say that an action is *productively successful* when it efficiently promotes the end it is designed to achieve, whereas an action is *constitutively successful* when it fulfills the necessary and sufficient conditions for being an instance of the type of activity it is supposed to be. Now, it is clear that the noncompliance of others can have the effect of undermining the productive success of your actions. This is true, moreover, whether or not the

[12] By "action," I mean to refer not only to external acts but to deliberative and volitional activity as well. In general it is important, when using the practical model, to guard against an overly legalistic construal of practice rules. It is clear, for example, that Rawls takes the rules of judicial punishment to constrain not only a judge's external actions, but also his internal actions of deliberation and choice. I believe this is because Rawls takes the actions falling under practice rules to constitute forms of active *participation* in the practice. I argue for this in my paper, "Three Conceptions of Action in Moral Theory," *Nous*, xxxv, 1 (2001): 93–117, and I make that argument in a different way below, section II.3.

end to be achieved is practice-dependent or practice-independent. That is, the conduct of others can undermine your productive success by interfering with your ability to promote ends that you must adopt in virtue of being a participant in the practice, as well as ends you adopt independently of that role. In this case, your interlocutor's noncompliance makes it quite unlikely that you will achieve your practice-dependent end of coming to a negotiated agreement, and he will also probably make it unlikely that you will achieve your practice-independent end of protecting the interests that are at stake in the dispute. But such productive failure is certainly compatible with constitutive success.[15] Indeed, in some cases it will be appropriate to describe the two of you as engaging in negotiation while still failing to reach an agreement or to protect the relevant interests. The question at hand is whether it is possible for the noncompliance of one party, at least beyond a certain threshold, to make it the case that the other is not even participating in the activity defined by the practice rules, despite the fact that the latter party is following the letter of the law.

In one sense, it is trivial that such constitutive interdependence exists. Every action, whether governed by a practice or not, depends to some extent upon the presence of a constitutive background if it is to be described in one way or another, and to the extent that such background conditions make reference to the behavior of others, our actions depend constitutively on those of others. For example, in order for my action to be describable as "giving you a hard time," you have to be annoyed or inconvenienced by what I am doing. But the absence of such conditions, it might be objected, does not always amount to a constitutive *failure* of any kind. For this to be the case, there has to be something my action *ought* to be, some character it is supposed to have but it lacks. In the negotiation example, it is only the case that you are *merely* babbling on with reference to some standard according to which you *ought* to have been negotiating. But, the objection goes, in this example you are simply mistaken about which rules are in effect. You are attempting to play the "negotiation game," while your interlocutor is attempting to play, say, the "domina-

[15] Whether the converse is true depends upon whether the end in question is practice-dependent or practice-independent. It is certainly not possible to win a baseball game if you are not playing, and in general if the end is a practice-dependent one like "winning," then constitutive success is a condition of productive success. But if the end is conceived as one for the sake of which an agent chose to enter into the practice—for example, to get some exercise—then productive success may not depend upon constitutive success. One might try to play baseball with a crew of five-year olds and end up getting a lot of exercise even if the group never succeeds in really playing the game.

tion game." As such, there is no established practice, and neither you nor he can be described as subject to the rules of negotiation in the first place.

In some cases this will indeed be an appropriate description of the situation. But the possibility I am inquiring into is one that presupposes that the practice is sufficiently well established (perhaps in part by means of external, legalistic criteria, but perhaps by means of implicit and explicit signs of commitment communicated by each party) for you and those with whom you interact to count as occupying roles in it and as being subject to its rules. In these cases, the noncompliant party is tacitly or implicitly claiming the protections and prerogatives attached to his role while at the same time failing to live up to its demands. If, under these conditions, it is *in principle* possible for such noncompliance to make it the case that you are babbling on rather than negotiating, this would show that your action depends on his, not only to be describable in one way or another, but to be describable in the way it *ought* to be described.

I believe the possibility of this sort of constitutive failure is presupposed by our ordinary use of the notion of a "sham." "This negotiation is a sham," you might say, pounding your fist on the table. To make this claim is to insist that the activity in which you are engaged is no longer what it ought to be, and that this has come about through no fault of your own, but rather through the misconduct of others.[14] Moreover you would normally make this charge as a way of calling into question the value of continuing to conform to the practice rules, given that your rule-governed actions have to some extent lost their integrity. What would make this sort of challenge intelligible? Given the constitutive nature of practice rules, the idea would have to be that conformity with the practice rules is no longer sufficient to insure constitutive success; that is, the fact that your action is chosen in accordance with the practice rules is no longer sufficient to make it count as participation in the form of activity that is, ideally, defined by those rules. Hence, when a practice has become a sham, adherence to the formal procedures in terms of which the activity is defined can take on the character of a mere formality, an empty gesture.

Aristotle wrote that a finger, when detached from the living organism of which it was a part, becomes a finger "in name only."[15] Similarly, I want to claim, lawfulness, when detached from the background

[14] Or through some other force external to your will. Here I am focusing on noncompliance as a potential source of nonideality, but as I mention in note 5, there are others.

[15] *Metaphysics*, 1035b24–25.

conditions in virtue of which its constitutive function is defined, can become merely nominal lawfulness—a kind of rectitude that is lacking, at least to some degree, in real value. I am going to call this the *sound and fury problem*, because it picks out a way in which forces beyond our control can deprive our well-chosen conduct of its proper significance. Now notice that if the sound and fury problem is a real possibility, it provides a model of what Nagel calls a "blind alley." As I mentioned earlier, a "blind alley" is a situation in which an ideally virtuous agent is somehow forced to act without justification. Now, if you are faced with the sound and fury problem, you are by hypothesis the occupant of a certain role, which means you are bound to govern yourself in accordance with a certain system of constitutive rules. But given that the practice is a sham, by so governing yourself, you end up engaging in constitutively failed action. If Rawls is right in claiming that the justification of an action falling under a practice can only appeal to the practice rules, then this constitutive failure amounts to a justificatory failure as well. For when you appeal to the practice rules for justification, you are attempting to justify your action as a form of participation in the activity defined by the practice. But when what you actually do amounts to something less than this—when, instead of making the moves, you end up merely going through the motions—then the practice rules cannot justify what you actually end up doing. Can you therefore appeal to some practice-independent standard in order to justify what you actually do? Not if, as was assumed, you are correct in taking yourself to occupy the role in question, and are not simply imagining yourself to do so. Unless you decide to abandon that role (which in some cases may not even be an option), you are bound by its rules. But unless your coparticipants change their ways, those rules will be constitutively and justificatorily defective.

To summarize: because you are a participant, you have to play, and because the rules are constitutive of participation, the only way to play is to play by the rules. But because the background conditions presupposed by such rules are ill established, playing by the rules fails to amount to participation in the relevant form of activity. As such, what is a conscientious player to do? If you comply with the letter of the law, you will betray its spirit, in the sense that you will not be engaging in the form of activity in terms of which you value yourself and your conduct as a player. If you violate the letter of that law, however, you will likewise fail to participate in that form of activity, because there are no other rules in terms of which that activity is defined.

II.3. How a practice becomes a sham. At this point an objection arises. The problem as I have just described it presupposes that there can

be a gap between adherence to the letter of the practice rules and full-blooded participation in the relevant form of activity. But if the rules are truly constitutive of the activity, how can any such gap exist? The idea that there can be such a gap seems to presuppose that we have an independent idea of what the activity ought to be like. But where does this idea come from? Without an answer to this question, the account I have offered is vulnerable to Nagel's suspicion that the standard to which we (perhaps unconsciously) appeal when we call a given practice a sham is none other than the principle of utility.[16] That is, we call a practice a sham when it has been rendered inefficient, or sufficiently inefficient, in achieving the practice-independent ends it was originally designed to promote. Above that threshold, we can afford to act like purists, abiding by the rules regardless of the consequences. But below that threshold, such purism becomes too costly, and we are justified in acting more pragmatically. And even if it is the case that when we act more pragmatically our actions have to be described somewhat differently—not as moves but as "shmoves"—so what? The point is to adopt the most efficient method available, given the circumstances, to advance certain practice-independent ends. Indeed, the efficiency of the practice is ultimately what justifies our taking up roles within it, so its inefficiency can justify our abandoning those rules in particular cases where adhering to them would be too costly.

Of course, this is the sort of position to which Rawls was originally responding in "Two Concepts of Rules," namely, a utilitarian account of practices that allows participants to take exception to practice rules on the basis of direct appeal to the principle of utility. The objection reveals that Rawls's position was not sufficiently defended, and to the extent that it was not, my own use of Rawls is likewise vulnerable. Here is what I think Rawls needs to complete the argument. Although it is the case that practices as wholes are adopted in virtue of their fitness to promote certain practice-independent ends, this efficacy is not the source of their normativity *as practices*, or of their bindingness upon us *as participants in practices*. The relevant contrast here is between a practice and a method. A method, as I think Rawls should conceive it, is a procedure that works like a tool; it is designed to serve some end, and that end is conceived as having been set in advance by a will that is constituted independently of the procedure. A practice, on the other hand, is a procedure by which a new, shared

[16] The same suspicion could of course be shared by a thoroughgoing utilitarian like Henry Sidgwick

will, along with its ends, is constituted.[17] This might sound counterintuitive, because as Rawls pointed out, we adopt practices as wholes for the sake of achieving certain practice-independent ends, and not obviously for the sake of constituting shared wills. Even if the initial incentive to set up a practice comes from the need to achieve a practice-independent end, however, the decision to achieve that end by means of a *practice* rather than a method amounts to a decision to claim that end as a common end, and to come together as a community in order to realize it. Hence we might say that, whereas a method defines a way of solving a problem, only a practice defines a way of making a problem, along with its solution, count as *ours*.

I think this idea underlies Rawls's account of the logical structure of practices. Rawls says that a practice is "the specification of a new form of activity."[18] But he is likewise committed to the view that a practice specifies a new source of agency, one to which that new form of activity is attributable. To see this, notice that actions falling under practices are attributable to agents not as particular individuals, but as occupants of roles or positions that are defined relationally in terms of one another. Such actions are thus attributable to individual agents only insofar as those agents are regarded as constituents of a shared source of agency. In this sense, the actions of each are the actions of all. They are essentially public actions, expressions of something like a shared will.[19] If actions governed by practice rules are to be valued only as public actions, then it would make sense to say that they are only properly described as such, in terms of the rules that constitute their public character.

Of course, this notion of a shared will is one that requires elucidation, and what I have to offer here is admittedly incomplete. There

[17] For a closely related distinction, see Anthony Laden's instructive discussion of the rational and the reasonable, in his *Reasonably Radical* (Ithaca: Cornell, 2001), chapter 4.

[18] "Two Concepts of Rules," p 36.

[19] This may seem counterintuitive in cases where the practice involves direct competition between opposing players or teams. When a pitcher for the Red Sox strikes out a batter for the Yankees, how is this action in any way attributable to the agency of all the players involved in the game? The answer is that every time a player acts within his role as a baseball player in a particular situation (for example, occupying a particular position relative to other players), he upholds the establishment of the practice. As such he contributes to the constitution of the practice-based activities of all involved. The pitcher for the Red Sox may thwart the Yankee batter's attempt to promote the specific practice-dependent end assigned to him in virtue of his position as a batter; but insofar as he acts well as a pitcher, he upholds the Yankee batter's status as a baseball player engaged in the practice of baseball. Another way to see this is to notice that when the competition takes place within a fundamentally cooperative context, playing to win is conceived as an aim that is fully compatible with good sportsmanship.

is a growing body of literature on the subject of shared agency, within which there is debate about what needs to be the case in order for it to be true that "we" are acting together. Some philosophers see the "we" as a structure of interlocking individual intentions to perform the activity, and some see it as a plural subject constituted by normative claims which individuals make on one another with respect to the performance of the activity.[20] For the purpose of this article, I need not take a stand on shared agency in general. However I am assuming, without argument, that at least in the specific case of a social practice, the "we" takes the latter form. In a practice, actions are attributable to a shared will because and insofar as participants make reciprocally-binding claims upon one another to adhere to the rules of the practice. This may make it look like I am employing a moralized notion of a practice, one which applies only to good or justifiable practices. The normative element being presupposed here is minimal, however. As I am understanding it, a practice must be characterized by reciprocity, but it need not be characterized by the highest degree of reciprocity. What is essential to a practice is that it is a system of social cooperation rather than a system of social manipulation; but this leaves room for practices that are less than fully democratic. For example, the institution of the Catholic Church defines a practice even though the decision-making structure of that institution is hierarchical. It is a practice because (or insofar as) it constitutes a system of social cooperation, rather than one based on coercion and force. It is cooperative in virtue of several features, which I put forward in a provisional way, and which are not intended to be exhaustive or conclusively formulated.[21] First, its rules are viewed by all participants not simply as commands backed by sanctions, but as genuine obligations. Moreover, participants view themselves and each other as responsible agents capable of following those rules on the basis of their own consciences. In addition, the system of rules is oriented towards the realization of a common good, a good the instantiation of which serves to benefit each member of the community. These conditions indicate the neighborhood in which the line between practices and nonpractices is to be drawn, though there will of course be borderline cases.[22]

[20] The first type of account is developed by Michael Bratman, *Faces of Intention* (New York: Cambridge, 1999), part II. The second type of account appears in the work of Margaret Gilbert. See her "Walking Together. A Paradigmatic Social Phenomenon," *Midwest Studies in Philosophy*, XV (1990): 1–14, and *On Social Facts* (Princeton: University Press, 1989)

[21] Here I am taking my lead from Rawls's discussion of a "decent consultation hierarchy" in his work on international relations, entitled *The Law of Peoples* (Cambridge: Harvard, 1999), see especially pp. 71–78.

[22] A further complication is that there will be cases in which an institution that has the internal structure of a practice relates to outsiders in a fundamentally uncoop-

The significance of the method/practice distinction is that it allows us to specify a nonconsequentialist basis for departing from the practice rules. For although the normativity of a practice *qua method* lies in its efficacy in promoting an end, its normativity *qua practice* lies in its fitness to constitute a shared will, a "we" to which the ends pursued, along with the moves themselves, are attributable. This is what the integrity of a practice consists in, and as long as it is intact, action in conformity with the rules will likewise have integrity as an expression of a public will, even when it fails to achieve its intended results. Under these circumstances, noncompliance is regarded as something accidental to the substance of the practice, something that therefore has no bearing on the stringency with which the practice rules bind participants. Now if all misconduct were properly regarded this way, then it would be impossible for noncompliance to render a practice a sham. That possibility rests on the assumption that at least some forms of misconduct can undermine what is essential to the practice, namely its constitutive function. Call this form of noncompliance *subversion* as opposed to mere *transgression*.[23] The concept of subversion, I take it, is that of a form of noncompliance that has the effect of detaching the law from the constitutive background in virtue of which lawful actions succeed in being expressions of a public will. The subversive is one who not only breaks but usurps the law, making it into a tool of his unilateral will. In doing so, he makes lawfulness into a mechanism by which his private will is served, rather than an ideal standard in virtue of which a public will is constituted and expressed.

The essential threat posed by subversion, then, is that it makes you end up serving a unilateral will, and not that it makes you end up promoting an end that is bad or evil.[24] Consider, for example, a

---

erative way. The example of a Mafia family, which I refer to below, fits this description I take it that Mafia institutions are sufficiently cooperative, at least internally, to constitute practices But it is an interesting question (and one I leave open here) whether a commitment to internal cooperation somehow leads to or implies a commitment to external cooperation.

[23] Korsgaard marks an analogous threshold between mere "imperfection" and "perversion" in "Taking the Law into Our Own Hands," p 318. She does not explain its basis, however.

[24] On the negative point, that what is at issue is not, fundamentally, the goodness or badness of your coparticipant's aim, my account is in agreement with the one Korsgaard gives in "The Right to Lie." However, (1) I do not claim at this point to have made an argument that applies directly to the case of practice-independent actions, like truthtelling, and (2) I believe my positive account of how you can be made a tool of evil, at least in the context of a practice, is more explicit than the account Korsgaard offers in that article. What I say here, however, is much more closely in accord with the position Korsgaard takes in her later article, "Taking the Law into Our Own Hands." Because that article deals with a problem concerning one's participation in a political institution, its relevance to the cases I am dealing

modification of the negotiation case in which your interlocutor is trying to promote the very practice-independent end in virtue of which you decided to engage in negotiation—namely that of protecting your interests. Imagine that your negotiating partner genuinely wants to find a resolution that adequately addresses your interests. But imagine that he does not want to present this resolution as the result of a process of negotiation, because he does not want to give the appearance that he has been responsive to your direct appeals. Such a person might purport to negotiate with you, while using the activity merely as a method for gathering information. Although he listens to what you say, he does not make a good faith attempt to come to a negotiated agreement with you on the basis of that interaction. His aim is to find out what you need and to develop a way of meeting your interests that he can later implement on his own, as a way, perhaps, of validating his conception of himself as a benevolent soul.

Despite the fact that the outcome at which this person is aiming is the same as the one you might have wanted to achieve through negotiation, my contention is that this sort of conduct can, at least at some point, render the negotiation a sham. For in this case your conegotiator is indeed using the activity as if it were a mere method, responding to you as a source of information but not as a source of reciprocally-binding claims. And this can put you in the same predicament of not being able to determine exactly what you are doing by acting in good faith in the face of such conduct.

This same point, that what makes noncompliance subversive is its unilateral nature rather than the value of its end, implies that the sound and fury problem can arise even in cases where the practice as a whole is oriented towards morally evil ends, and subversion is undertaken for the sake of the morally good. Consider a modification of the previous scenario. Suppose you are a middle-level Mafioso, and you have come to consult with the Godfather. The family code requires that you engage in consultation in this case—one in which, say, you need the Godfather's permission to take over a new market. Because the internal organization of the family is that of a practice, you come to the Godfather out of a sense of duty, and not simply out of fear

---

with here is obviously more direct. But there are other important differences between this later article and the earlier one. In the earlier one, for example, Korsgaard suggests that appeal to the Formula of Universal Law can give direct guidance to the person faced by the murderer at the door. In the revolution article, she denies that unriversalization is of any help to the revolutionary (p. 319). I believe my position, once it is extended to Kantian theory generally, will be in accord with Korsgaard's later view; I should emphasize again, however, that I do not claim to have yet made that extension.

that you will be punished if you do not. But now you notice a bump under his shirt, and you become aware, to your horror, that he is wearing a wire. Instead of consulting with you, he is using the consultation as a way of gathering evidence for the FBI. Perhaps his motive for doing this is self-interest; he has concluded that sooner or later he will be convicted, and has decided to cut a deal with the authorities so as to insure a lighter sentence. Or perhaps he has undergone a personal moral conversion, and he feels called upon to help the authorities bring his "family" members to justice for its own sake. In the former case, his noncompliant action is morally good, albeit only externally so. In the latter case, his action is morally good in the full sense of being the right action chosen for the right reason. From your point of view *as a Mafioso*, however, the moral status of his action makes no difference. In either case, the Godfather is acting on the basis of principles that are practice-independent, and to this extent he has constituted his will as a private one, while still attempting to make use of the public procedures of consultation. Insofar as you are committed to being a *good Mafioso*, his moral conviction operates as a subversive influence.

The type of case that might seem hardest to account for on the view I am proposing is one in which the practice is morally good and has integrity, but it fails to promote the practice-dependent ends that all participants, *qua* participants, are committed to realizing. To take a familiar example, let us suppose the judicial system is intact, but because of the inevitable imperfections inherent in any such system, a guilty person goes free, or an innocent person is condemned. A good participant will of course deeply regret the outcome and perhaps try to work toward improving the practice so as to make this sort of result less likely in the future. If the normativity of the practice, however, does stem from its constitutive function, and if that function is intact, this need to change the rules does not provide a justification for breaking them. That said, it might be argued that there must be some degree of inefficiency so egregious that it would justify bending the practice rules, at least where the end in question is practice-dependent.[25] On the account I am proposing, however, the egregiousness of the inefficiency is only relevant insofar as it serves as an indication that the public character of the procedure is not, in fact, intact. In cases of jury nullification, for example, it is the law's integrity as an expression of the general will that is at stake. And to establish that the integrity of that law has been undermined, one would have

[25] I again thank Shiffrin for prompting me to address this sort of objection.

to argue that the law has been made to serve an alien will. When this is the case, the egregiously bad outcome is to be avoided not because it is an imperfect realization of a public aim, but because it is the successful realization of a private one.

II.4. An illustration. To make my rather abstract claims more vivid, here is a more detailed illustration of the predicament I am calling a "blind alley," albeit one that involves subversion for the sake of a morally evil end. I offer it with the caveat that nothing in my argument turns on your agreeing that this particular case counts as one in which the practice rules have indeed been subverted. All I need is your agreement that the example is illustrative of the sort of features that would characterize such a condition. The example is taken from the film *L. A. Confidential*.[26] Among other things, this is a story about a rookie cop who vows to play clean despite the fact that everyone else plays dirty, only to find that playing clean amounts to violating the very ideal of justice he set out to uphold.[27] The climactic scene comes in a confrontation at the end between the cop, Edmund Exley, and the corrupt Chief of Police, Captain Dudley. Dudley has been using his influence as Chief of Police to gain control of a drug trade formerly run by the Mafia. In an effort to secure his power, he has framed innocent civilians and has tortured and murdered fellow police officers. Despite an extensive cover-up operation, Exley has discovered the truth, and Dudley stages an ambush in which he tries to have Exley murdered. After a bloody shootout, Exley ends up alone with Dudley, holding the Captain at gunpoint. Police sirens are heard approaching from the distance. Dudley says to Exley, "Well, kid—are you going to shoot me, or arrest me?" When Exley does not shoot, Dudley says, "Always the politician, eh? Good boy. Just let me do the talking and you'll be Chief of Detectives." The Captain walks out to face the approaching police cars, leaving the armed Exley behind him. Then he pulls out his badge and holds it up, announcing to Exley, "Hold up your badge, so that they know you're a police officer." At that point Exley shoots Dudley in the back. Then Exley goes out to face the cars, holding his badge in the air.

The scene poses the question vividly: in this situation, which course

[26] The film is based on the novel of the same name by James Ellroy (New York: Warner, 1990); film released by Warner Brothers, screenplay by Brian Helgeland and Curtis Hanson. The scene I describe occurs only in the film version.

[27] Actually, the motives of the clean cop are not quite that simple—at the outset of the story, he is driven more by a desire to establish a reputation as a clean cop than by a desire to be one. But I take it that in the scene I describe, which comes at the end, he acts on the latter motive

of action—shooting or arresting—is demanded by the practice of policing, and by Exley's ideal of himself as a good policeman? If there is any justification for Exley's decision to shoot, it cannot simply be that, in this case, it was likely that the Captain would have eluded prosecution. A good cop is not entitled to choose between arresting and shooting a suspected lawbreaker on the basis of an estimation of the likelihood that he will be brought to trial and convicted. This is so even if the crimes involved are egregious, and if the likelihood of his repeating the offense if allowed to go free is high. And this is so even in light of the fact that the whole purpose of establishing a practice of law enforcement is to provide a mechanism for restraining and punishing those who break laws. Granted, if law enforcement were merely a *method* for achieving this goal, Exley might be straightforwardly justified in doing whatever is necessary to stop Dudley. But since law enforcement is a practice, or perhaps a subpractice within the general practice of citizenship, acts of law enforcement must, first and foremost, be public acts; they must be undertaken in such a way as to represent the united will of the citizenry as a whole, a will that is embodied more specifically in the standards and procedures of policing. As such, it looks like a good cop has no choice but to make the arrest. For if he shoots, he acts unilaterally, on his own private will, and he becomes a murderer just like Dudley.

And yet we are supposed to see Exley as being faced with a real dilemma (and most of us, by the end of the story, do feel a sense of dilemma). How can this be explained? One option is to see Exley as being torn between his conception of himself as a good cop and some other conception of himself that carries with it an obligation to promote utility in general (or to prevent suffering in general). This is the Nagelian, pluralist view, according to which our sense of moral dilemma in extreme situations like this stems from the fact that we are torn between deontological and utilitarian perspectives. But this reading of the conflict seems strained. Exley is torn precisely because he feels obligated to do what is demanded of him *as* a good policeman; indeed after he shoots Dudley, he holds his badge in the air, as if to show that he is (still) a good policeman. Hence I take it that in the decision between arresting and shooting, what is at stake for him is his integrity as a police officer, independent of any further obligation to promote the overall good. But if this is indeed Exley's concern, then why is it not obvious that he ought to make the arrest? What could tempt a conscientious cop to violate the constitutive rules of policing?

The argument I have been developing suggests an alternative interpretation. It is open to Exley to ask himself whether the Captain's criminal activity is of a nature as to undermine the public character

100 J826

of the practice, and hence that of Exley's own lawful conduct. Has the Captain merely transgressed the law, or has he subverted it? The Captain's last words and gestures are, I take it, supposed to suggest the latter: "Hold up your badge, so that they know you're a police officer." The image suggests, albeit metaphorically, that the Captain is not merely misapplying the concept of a policeman to himself—he is trying to change the function of that concept, to make it stand for all he stands for—namely, the unconstrained pursuit of personal power.[28] This indicates that from Exley's point of view, the Captain is putting himself above the law in such a way as to *alter its constitutive function*, making it a tool of his unilateral will rather than a procedure by which a public will is constituted and expressed. If this is indeed how Exley views his situation, then he can rightly wonder whether making the arrest, though formally a public act, amounts substantively to anything more than submission to the Captain's private will. And by the same token, he can rightly wonder whether shooting the Captain, though formally an act of murder, might in some sense amount to a *surrogate* for the public act of law enforcement that he must—but is not in a position to—carry out.

My argument does not require that we agree that Exley is right in thinking that the situation he faces in fact meets this description. Moreover, even if we agree with his assessment, we need not agree that the particular form of lawbreaking he ultimately opts for—namely shooting the Captain—is the least unlawful of unlawful alternatives open to him.[29] What matters is that *if* Exley sees his situation the way

[28] Korsgaard argues that what Kant finds horrifying about one particular unjust act, the formal execution of a monarch, is that "it presents an evil act in the outward form of a lawful one," and is in this respect "like an act of a malevolent [diabolical] will"—"Taking the Law into Our Own Hands," p. 318. Captain Dudley's holding up of his police badge has the same property. But there is still a question whether being faced with this particular kind of horror gives us license to act in ways that would otherwise be disallowed. Korsgaard suggests an answer along the lines I am proposing here when she says, "[C]oncern for human rights leads the virtuous person to accept the authority of the law, but in such circumstances adherence to the law will lead her to support institutions that systematically violate human rights" (p. 318). I am trying to give a more detailed and more explicitly nonconsequentialist account of what this problem is, such that it can weaken the force with which the agent is bound by the letter of the law. On my account, it is only to the extent that the law has been made into the tool of another's will that you can, in virtue of your lawfulness, be made a tool of another's will.

[29] Although the account I have been offering here is independent of Kant's theory, it does seem worth mentioning Kant's insistence that he "cannot deny all respect to even a vicious man as a human being; I cannot withdraw at least the respect that belongs to him in his quality as a human being, even though by his deeds he makes himself unworthy of it"—*Metaphysics of Morals*, p. 580/6:463. On Kant's view, one can never become so wicked as to lose one's *nominal* status as a human being. Just what this claim ends up ruling in and out is a further question. Kant took this claim

KP 2257

I have described, he will indeed find himself in a predicament with the structure of a blind alley. As a good policeman, he is required to refrain from vigilantism; he must perform only those acts that can count as public acts, even at the cost of allowing a guilty and perhaps dangerous person to go free. And yet he takes the situation to be one in which the law itself is being held hostage; it is as if the Captain, through an act of moral ventriloquism, has substituted his own voice for the voice of the people, while leaving the outward procedures of public expression in place. If this is so, then Exley's adherence to those formal procedures is at risk of amounting to an empty formality.[30] But departure from them is not without risk as well. For those procedures are the only ones that have any claim at all to express the public will, so that departure from them is not, by any publicly-communicable standard, a way of embodying that will.

Let me be clear about what I am *not* claiming here. I am not claiming that every instance of noncompliance on the part of one or more parties renders a practice a sham. Practices are inevitably imperfect realizations of the ideals that govern them, and it is in general incumbent on us to abstract from these imperfections in conducting our ordinary affairs. What I am claiming, however, is that if noncompliance can *in principle* render a practice a sham, in the normatively significant sense that I am outlining here, this is how we have to conceive of it doing so. There is, no doubt, a threshold judgment that needs to be made in these cases, and what I am suggesting is that there is a nonconsequentialist standard to be appealed to in making that judgment. That standard is one which assesses the integrity of the practice as a practice, rather than its efficacy as a method, and hence it involves a judgment about how you would be relating yourself to the other participants in the practice were you to abide by the letter of a law that is being misused. It is entirely possible that the threshold set by this standard turns out to be extremely high.

---

to justify a prohibition against dehumanizing forms of punishment, and yet he was a proponent of capital punishment in certain cases. Presumably preservation of one's nominal status as a human being does not rule out the possibility that the claims one makes on others in virtue of that status might be weakened, if one has acted in a way that is essentially incompatible with that status. In such a case, we might say that one's status as a human being has become *merely* nominal.

[30] This is not to deny that there can sometimes be value in adhering to a formal procedure even when it has, to some extent, become an empty formality. One might adhere to the law as an "aspirational" gesture, a way of inviting the other back into the moral community, and as an expression of hope that the proper function of the law might be restored. My view implies only that to adhere to the law in this way is somewhat different from adhering to it under conditions where its constitutive function is unimpaired. Thanks to Shiffrin for prompting me to address this.

It may be that there are very few cases in which noncompliance has this *subversive* character, in which case there would very seldom be grounds for departing from strict adherence to the letter of the practice rules. But to deny that any such threshold exists in principle, I maintain, is tantamount to claiming that participation in a practice is a kind of blind proceduralism. If what is valued is adherence to the letter of the practice rules as such, independent of any deeper conception of what one is thereby doing, then practice rules might indeed be *incorruptible*. Noncompliance would not threaten their constitutive function, because they would not have one. But the price of such incorruptibility would be to render them pointless. The point of setting up a practice is to establish a way to act *as one* in relation to some problem or other. The normativity of practice rules depends upon their playing this constitutive, unifying role. Hence, we can at least conceive of those rules losing their normative grip should they be forced to serve some alien function. The noncompliance of participants is at least one potentially corrupting influence on a practice—one type of nonideal condition—though there may be others.[31] What is important is that this construal of what nonideality consists in provides a nonconsequentialist basis for the claim that the wrongdoing of others can, at least within the context of a social practice, mitigate the stringency with which otherwise unexceptionable rules apply to us.

III. IMPLICATIONS

III.1. How this account sheds light on the possibility of moral dilemmas. This analysis shows, first of all, that a blind alley can arise without supposing that the agent is torn between his practice-dependent obligations on the one hand and utilitarian considerations on the other. It arises simply in virtue of his commitment to the practice, combined with the fact that the subversive activity of another has made the practice constitutively defective. But this account can also explain why action under nonideal conditions can appear to be more goal oriented than action under ideal conditions, and hence why it can look superficially like action governed by consequentialist reasoning. The explanation is as follows. Participation in a practice is not blind rule-following; it is a self-conscious form of activity in which an agent sees himself as

[31] Here I am referring to the other main part of nonideal theory as it is described by Rawls, the part which deals with problems arising out of "natural limitations and historical contingencies," as opposed to those arising out of the noncompliance of others—see *A Theory of Justice*, p. 246. I have tried to argue that human immaturity constitutes one kind of nonideal condition in my "What Is a Child?" *Ethics*, CIX (July 1999): 715–38.

relating to others as a coparticipant in virtue of his adherence to the rules. As such, every participant must be conceived as adhering to the rules on the basis of a background assessment of the integrity of the practice as a whole. Under nonideal conditions, this background assessment comes to the fore, because the practice is judged to be in need of repair. This can seem to force the participant into the position of the social engineer, since the social engineer is one who takes the functioning of the practice as a whole as his object. But the similarity is superficial. To distinguish the standpoint of the social engineer from the standpoint assumed by a participant under nonideal conditions, I will call the latter the perspective of the *reformer*.[32] The reformer differs from the social engineer in two crucial respects: (1) the reformer takes himself to be a participant in the practice, and hence to be bound by the practice (construed broadly) in every action, whereas the social engineer, on the other hand, does not think of himself as a participant; and (2) the reformer adopts the end of bringing the practice into conformity with its internal constitutive standard, whereas the social engineer attempts to improve its efficiency relative to an external standard.

As for (1), because a reformer is simply a participant under nonideal conditions, a reformer is first and foremost committed to acting as far as possible in accordance with the demands of the practice, albeit in a situation where this may or may not mean acting according to the "letter" of its law. It is only provided this procedural condition is fulfilled that the reformer is entitled to work toward bringing about any specific state of affairs. As for (2), the reformer is a participant who, in the face of nonideal conditions, may find it necessary (for the sake of preserving the integrity of his own action) to adopt the end of repairing the practice as a whole, or at least of not acting in such a way as to undermine that end. But his aim here is not to make the practice more efficient in producing results that are valued independently of the practice; rather, his aim is to restore the integrity of the practice as a practice, to bring about the conditions under which it would count as the genuinely shared form of activity it purports to

[32] Rawls comes close to making this distinction on pp. 32–33 of "Two Concepts of Rules," where he writes, "...[A]s with any set of rules there is understood a background of circumstances under which it is expected to be applied and which need not—indeed which cannot—be fully stated Should these circumstances change, then even if there is no rule which provides for the case, it may still be in accordance with the practice that one be released from one's obligation But this sort of defense allowed by a practice must not be confused with the general option to weigh each particular case on utilitarian grounds which critics of utilitarianism have thought it necessarily to involve." This paper can be read as an attempt to interpret this paragraph, and so to shed light on Rawls's later notion of nonideal conditions.

be. Thus, the existence of a superficial similarity between the standpoints of the social engineer and the reformer helps to explain why action under nonideal conditions can appear to be governed by consequentialist reasoning; but further examination of those standpoints reveals this appearance to be misleading.[33]

An additional virtue of my account is that it avoids glib moralism, by showing how a moral dilemma can arise through no cognitive or volitional failure on the part of the agent. The predicament I have described is one that retains its dilemmatic character even when the agent is fully aware of its source and structure. And it does so because the situation is, at least in one sense, one in which there is no fully justified course of action to be taken. The agent has to perform a public action, because to do anything less would amount to failure to act in his public role, a role in terms of which he values himself and his conduct. But the situation is such that none of the alternatives straightforwardly meet that standard. Each of them fails, either formally or substantively, to be the public action it ought to be. In this sense, there is no right thing to do. I take it this way of looking at the situation provides grounds for the rather puzzling, retrospective feeling which Nagel calls "agent-regret."[34] *Agent-regret* is the feeling that you have violated a standard that somehow both applies to you under the circumstances and is, given the circumstances, impossible to live up to. On this account, your role as a participant sets a standard of conduct that applies to you whether or not any of the alternatives actually open to you meet that standard.

My account is at bottom, however, ambivalent about agent-regret. For there is another way of looking at the situation in which there is a right thing for you to do—namely, to do the best you can to give your action a public character, by choosing the least defective from a set of defective moves. That might turn out to be an action in conformity with the letter of the law, or it might not. But as long as you choose the surrogate that most closely approximates the governing ideal, you will have done the best anyone in your position could have done. This way of looking at the situation provides grounds for thinking that you are free of wrongdoing in any straightforward sense.

[33] As Tony Laden has pointed out to me, this analysis shows why it is good to build mechanisms of reform into the structure of practices. Checks and balances which allow authorities at higher levels to repair damage done at lower levels (for example, judicial review, constitutional amendment, and so forth) allow participants to change the rules while still playing by them. In this way such mechanisms narrow the gap between reformers and social engineers.

[34] Nagel coins this term in "Moral Luck," in his collection *Mortal Questions*, pp. 24–38, see p. 37.

On the account I have offered, this fundamental ambivalence about agent-regret is rooted in the fact that the reformer's unlawful actions can be regarded as differing from lawful actions either in kind or in degree. They differ in kind from lawful actions insofar as they are actions undertaken independently of the practice rules, and hence on the authorization of nothing other than a unilateral will. On this way of looking at such actions, they are simply lacking in public character, and in this respect are no different from the actions of an outlaw. From this perspective, agent-regret may be warranted. But insofar as such actions are regarded as surrogates for lawful actions, undertaken out of your conscientious sense of what the spirit of the law requires in this unusual circumstance, then such actions may be regarded as approximations of lawful conduct, so that the difference between the former and the latter is one of degree. From this perspective agent-regret is not warranted.

Is this ambivalence a virtue or a vice of my account? One reason for thinking it is a virtue is that our ordinary reactive attitudes display the same kind of ambivalence. When an agent has chosen the lesser of evil courses of action open to him under conditions of corruption, we tend to think, on the one hand, that it is inappropriate to hold him responsible for any wrongdoing. Yet we also tend to think that it is at least *not inappropriate* for the agent, reflecting on his own conduct, to feel that he has somehow compromised his integrity, or that he has been left with "dirty hands." A theory that unambivalently endorsed or rejected agent-regret would have trouble accounting for one or the other of these intuitions. I believe my account provides a basis for both, without falling into incoherence.[36]

III.2. Implications for Kantian nonideal theory. I have explained the implications of the view for the problem of moral dilemmas; at least insofar as such dilemmas can arise within the context of social practices. But my primary and longer-term interest is in relating the view I have laid out to Kant's moral theory in such a way as to show how the scope of Kant's purism might be limited. This obviously requires a separate article, but the basis of a strategy for doing this has, I think, been put in place. What I believe I have shown is that rules that bind *categorically* are not for that reason *incorruptible*, at least when their categorical status depends upon their serving a constitutive function.

[36] Given what I have said, this implies that the agent can take the "difference-in-kind" view of his unlawful action, whereas the observer ought to take the "difference-in-degree" view. My account would be strengthened if I could show why it makes sense for the points of view to line up in this way, but at present I do not have such an explanation.

Practice rules can serve as models of categorical imperatives, because they apply unconditionally to all participants. If you are in the role of a participant, then you must play, because that is the only way to act *as* a participant. And the only way to play is to play by the rules, because the rules are (at least ideally) constitutive of participation.[36]

Now, practice rules are only models of categorical imperatives, because normally we have the option of abandoning any conventionally-established practice. But what if there were a non-optional practice, a role we had to occupy on pain of not being agents at all, and a set of rules we had to follow on pain of not doing anything at all? If abandoning agency is not a human possibility, then those rules would in the fullest sense be categorical imperatives, and we would be bound to adhere to them under all conditions. But if those rules were categorically binding in virtue of their constitutive role—in virtue of their fitness to constitute a shared will and a shared form of activity attributable to that will—then they would by the same token be vulnerable to corruption; in principle, there would be some threshold below which noncompliance could force the letter and the spirit of the moral law itself to come apart. In this unfortunate circumstance, the conscientious agent would be forced to ask whether his *regard for the universal form* of his maxim had itself been rendered something of a *mere formality*.

I have not yet argued for this claim, in large part because many of the actions required by the moral law, for example truthtelling, are not obviously dependent upon anything like a practice to be what they ought to be. A separate argument is thus needed to show how the extension of the argument offered here to such cases as that of the murderer-at-the-door is possible. The argument would have to claim that what is required by the moral law, strictly speaking, is not truthtelling *per se* but some mode of person-to-person, or, in Kant's language, legislator-to-legislator relation. In this paper I claim only to have shown that within a practice, there is indeed a threshold beyond which others' wrongdoing can make your compliance count as a form of complicity.

TAMAR SCHAPIRO

Stanford University

[36] The strategy of accounting for categorical normative force by appeal to constitutive function has been developed to some extent by Korsgaard, especially in her article, "Self-Constitution in the Ethics of Plato and Kant," *Journal of Ethics*, III (1999): 1–29. As Korsgaard mentions in that paper, a challenge facing any defender of this strategy is to explain how defective action is possible; how an action which violates a constitutive rule can count as failing to be what it ought to be, instead of simply counting as a successful instantiation of something else. I discuss this in relation to practice rules in my "Three Conceptions of Action in Moral Theory."

SOME PROBLEMS FOR CONDITIONALIZATION AND REFLECTION*

I will present five puzzles that show that rational people can update their degrees of belief in manners that violate Bayesian conditionalization and Bas van Fraassen's reflection principle. I will then argue that these violations of conditionalization and reflection are due to the fact that there are two as yet unrecognized ways in which the degrees of belief of rational people can develop.

I. TWO ROADS TO SHANGRI LA

Every now and then, the guardians to Shangri La will allow a mere mortal to enter that hallowed ground. You have been chosen because you are a fan of the Los Angeles Clippers. But there is an ancient law about entry into Shangri La: you are only allowed to enter, if, once you have entered, you no longer know by what path you entered. Together with the guardians, you have devised a plan that satisfies this law. There are two paths to Shangri La, the Path by the Mountains, and the Path by the Sea. A fair coin will be tossed by the guardians to determine which path you will take: if heads you go by the Mountains, if tails you go by the Sea. If you go by the Mountains, nothing strange will happen: while traveling you will see the glorious Mountains, and even after you enter Shangri La, you will forever retain your memories of that Magnificent Journey. If you go by the Sea, you will revel in the Beauty of the Misty Ocean. But, just as you enter Shangri La, your memory of this Beauteous Journey will be erased and be replaced by a memory of the Journey by the Mountains.

Suppose that in fact you travel by the Mountains. How will your degrees of belief develop? Before you set out your degree of belief in heads will be ½. Then, as you travel along the Mountains and you gaze upon them, your degree of belief in heads will be one. But then, once you have arrived, you will revert to having degree of belief ½ in heads. For you will know that you would have had the memories that you have either way, and hence you know that the only relevant information that you have is that the coin was fair.

This seems a bizarre development of degrees of belief. For as you

*I would like to thank John Collins, Adam Elga, John Hawthorne, Isaac Levi, Barry Loewer, and Tim Maudlin for extensive and crucial comments and discussions on earlier versions of this article.

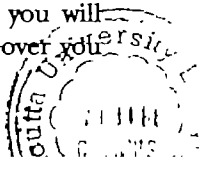
0022-362X/03/0007/356–70 © 2003 The Journal of Philosophy, Inc.

are traveling along the Mountains, you know that your degree of belief in heads is going to go down from one to ½. You do not have the least inclination to trust those future degrees of belief. Those future degrees of belief will not arise because you will acquire any evidence, at least not in any straightforward sense of "acquiring evidence." Nonetheless, you think you will behave in a fully rational manner when you acquire those future degrees of belief. Moreover, you know that the development of your memories will be completely normal. It is only because something strange would have happened to your memories had the coin landed tails that you are compelled to change your degree of belief to ½ when that counterfactual possibility would have occurred.

II. THE PRISONER

You have just been returned to your cell on death row, after your last supper. You are to be executed tomorrow. You have made a last minute appeal to President George W. Bush for clemency. Since Dick Cheney is in the hospital and cannot be consulted, George W. will decide by flipping a coin: heads you die, tails you live. His decision will be made known to the prison staff before midnight. You are friends with the prison officer who will take over the guard of your cell at midnight. He is not allowed to talk to you, but he will tell you of Bush's decision by switching the light in your cell off at the stroke of midnight if it was heads. He will leave it on if it was tails. Unfortunately you do not have a clock or a watch. All you know is that it is now 6 PM since that is when prisoners are returned to their cells after supper. You start to reminisce and think fondly of your previous career as a Bayesian. You suddenly get excited when you notice that there is going to be something funny about the development of your degrees of belief. Like anybody else, you do not have a perfect internal clock. At the moment you are certain that it is 6 PM, but as time passes your degrees of belief are going to be spread out over a range of times. What rules should such developments satisfy?

Let us start on this problem by focusing on one particularly puzzling feature of such developments. When in fact it is just before midnight, say 11:59 PM, you are going to have a certain, nonzero, degree of belief that it is now later than midnight. Of course, at 11:59 PM the light in your cell is still going to be on. Given that at this time you will have a nonzero degree of belief that it is after midnight, and given that in fact you will see that the light is still on, you will presumably take it that the light provides some evidence that the outcome was tails. Indeed, it seems clear that as it gets closer to midnight, you will monotonically increase your degree of belief in tails. Moreover, you



know in advance that this will happen. This seems puzzling. Of course, after midnight, your degree of belief in tails will either keep on increasing, or it will flip to zero at midnight and stay there after midnight. But that does not diminish the puzzlement about the predictable and inevitable increase in your degree of belief in tails prior to midnight. In fact, it seems that this increase is not merely puzzling, it seems patently irrational. For since this increase is entirely predictable, surely you could be made to lose money in a sequence of bets. At 6 PM you will be willing to accept a bet on heads at even odds, and at 11:59 PM you will, almost certainly, be willing to accept a bet on tails at worse than even odds. And that adds up to a sure loss. And surely that means you are irrational.

Now, one might think that this last argument shows that your degree of belief in tails in fact should not go up prior to midnight. One might indeed claim that since your degree of belief in heads should remain ½ until midnight, you should adjust your idea of what time it is when you see that the light is still on, rather than adjust your degree of belief in tails as time passes. But of course, this suggestion is impossible to carry out. Armed with an imperfect internal clock, you simply cannot make sure that your degree of belief in heads stays ½ until midnight, while allowing it to go down after midnight. So how should they develop?

Let us start with a much simpler case. Let us suppose that there is no coin toss and no light switching (and that you know this). You go into your cell at 6 PM. As time goes by there will be some development of your degrees of belief as to what time it is. Let us suppose that your degrees of belief in possible times develop as pictured in the top half of Figure 1.

Next, let us ask how your degrees of belief should develop were you to know with certainty that the guard will switch the light off at midnight, 12 PM. It should be clear then that at 11:59 PM your degree of belief distribution should be entirely confined to the left of midnight, as depicted in the bottom half of Figure 1. For at 11:59 PM the light will still be on, so that you know that it must be before 12 PM. But other than that it should be entirely confined to the left of 12 PM, it is not immediately clear exactly what your degree of belief distribution should be at 11:59 PM. It is not even obvious that there should be a unique answer to this question. A very simple consideration, however, leads to a unique answer.

Suppose that, even though the guard is going to switch the light off at 12 PM, you were not told that the guard is going to switch the light off at 12 PM. Then the development of your degree of belief would be as pictured in the top half of Figure 1. Next, suppose that

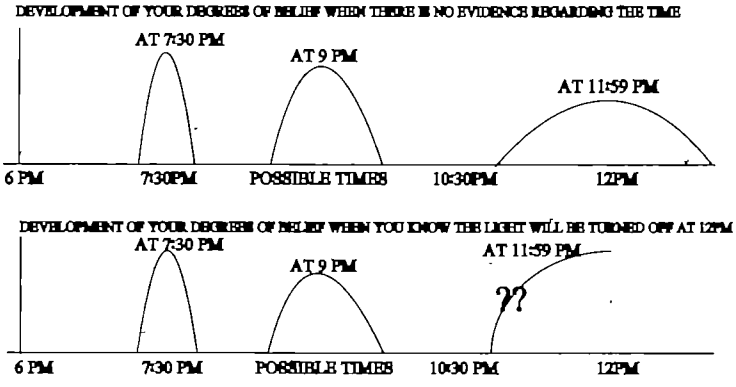


Figure 1

at 11:59 PM you are told that the guard will switch the light off at 12 PM, but you are not told that it is now 11:59 PM. Obviously, since the light is still on you can infer that it is prior to 12 PM. Surely you should update your degrees of belief by conditionalization: you should erase that part of your degree of belief distribution that is to the right of 12 PM, and renormalize the remaining part (increase the remaining part proportionally). Now it is clear that this is also the degree of belief distribution that you should have arrived at had you known all along that the guard would turn the light off at 12 PM. For either way you have accumulated exactly the same relevant information and experience by 11:59 PM. This uniquely determines how your degree of belief distribution should develop when you know all along that the guard will turn the light off at 12 PM. At any time this (constrained) distribution should be the distribution that you arrive at by conditionalizing the distribution that you have if you have no evidence regarding the time, on the fact that it is now before 12 PM. One can picture this development in the following way. One takes the development of the top part of Figure 1. As this distribution starts to pass through the 12 PM boundary, the part that passes through this boundary gets erased, and, in a continuous manner, it gets proportionally added to the part that is to the left of the 12 PM boundary.

Now we are ready to solve the original puzzle. Your degrees of belief in that case can be pictured as being distributed over possible times in two possible worlds: see Figure 2. The development is now such that when the bottom part of the degree of belief distribution hits midnight, it gets snuffed out to the right of midnight, and the rest of the degree of belief distribution is continuously renormalized, that is, the top part of the degree of belief distribution and the

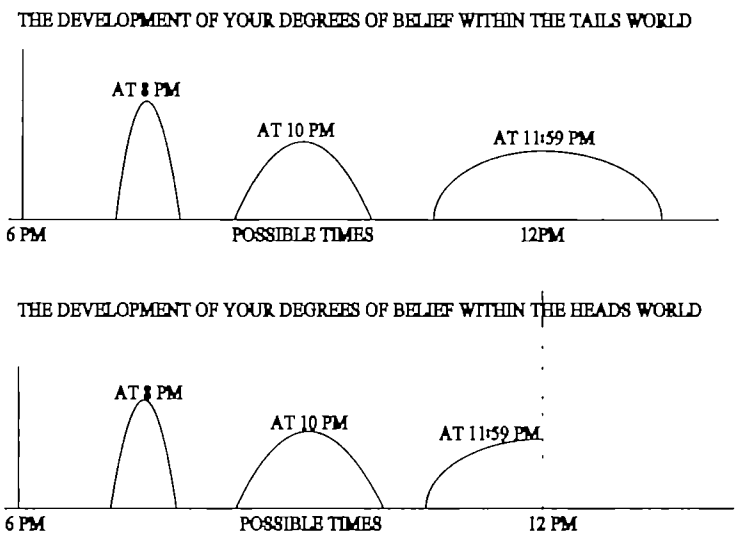


Figure 2

remaining bottom part are continuously proportionally increased as time passes. Note that Figure 2 is essentially different from Figure 1. In Figure 2, the top distribution starts to increase its absolute size once the leading edge of the bottom distribution hits midnight. This does not happen in Figure 1, since there the degree of belief distributions each were total degree of belief distributions in separate scenarios. Also, in Figure 2, the bottom distribution starts to increase in size once its leading edge hits midnight, but it only increases half as much as it does in Figure 1, since half of the "gains" is being diverted to the top degree of belief distribution.

Thus, at the very least, until it actually is midnight, the top and the bottom degrees of belief distribution will always be identical to each other, in terms of shape and size, to the left of midnight. Prior to midnight, your degrees of belief will be such that conditional upon it being prior to midnight, it is equally likely to be heads as tails. Your unconditional degree of belief in tails, however, will increase monotonically as you approach midnight.

After midnight there are two possible ways in which your degree of belief distribution can develop. If the light is switched off, your degree of belief distribution collapses completely onto midnight and onto the heads world. If in fact it is not switched off, your degree of belief distribution continues to move to the right in both worlds, and it continues to be snuffed out in the heads world to the right of

midnight, and the remaining degrees of belief keep being proportionally increased.[1]

Now I can answer the questions that I started with. It is true, as I surmised, that your degree of belief in tails will have increased by 11:59 PM. You will take your internal sense of the passing of time, and combine it with the fact that the light is still on, and you will take this as providing some evidence that the outcome is tails. It is also true, as I surmised, that the light still being on will be taken by you as providing some evidence that it is not yet midnight. For at 11:59 PM your degree of belief distribution over possible times (averaged over the heads and tails worlds) will be further to the left than it would have been had you believed that the light would stay on no matter what. More generally, we have found a unique solution to the puzzle of how a rational person's sense of time must interact with evidence, given how that person's sense of time works in the absence of evidence.

Rather surprisingly, this interaction can be such, as it is in my example, that you know in advance that at some specified later time you will, almost certainly, have increased your degree of belief in tails, and that you could not possibly have decreased your degree of belief in tails.[2] It is also interesting to note that nothing essential changes in this example if one assumes that the coin toss will take place exactly at midnight. Thus it can be the case that one knows in advance that one will increase one's degrees of belief that a coin toss, *which is yet*

[1] Thus, for instance, if the light is not switched off, there must be a moment (which could be before or after midnight) such that you have an equal degree of belief in each of the three possibilities: heads and it is before midnight, tails and it is before midnight, tails and it is after midnight.

[2] One might wonder why I inserted the phrase 'almost certainly' in this sentence The reason for this is that there is a subtlety as to whether you know at 6 PM that you will have an increased degree of belief in tails at 11:59 PM. There is an incoherence in assuming that at 6 PM you know with certainty what your degree of belief distribution over possible times will be at 11:59 PM. For if you knew that, you could simply wait until your degree of belief distribution was exactly like that. (You can presumably establish by introspection what your degree of belief distribution is.) And when you reach that distribution, you would know that it has to be 11:59 PM. So when that happens you should then collapse your degree of belief distribution completely on it being 11:59 PM. But this is incoherent. Thus, the fact that you do not have a perfect internal clock also implies that you cannot know in advance what your degree of belief distribution is going to look like after it has developed (guided only by your internal clock). Thus you cannot in advance be certain how your degree of belief distribution over possible times will develop. Nonetheless, you can be certain at 6 PM that your degree of belief in tails will not decrease prior to midnight, and that it is extremely likely to have increased by 11:59 PM. At 6 PM your expectation for your degree of belief in tails at 11:59 PM will be substantially greater than ½

to occur, will land tails. Of course, at the time that one has this increased degree of belief one does not know that this coin toss is yet to occur. Nonetheless, such predictable increases in degrees of belief seem very strange.

III. JOHN COLLINS'S PRISONER

John Collins has come up with the following variation of the case of the prisoner that was described in the previous section. In Collins's variation, the prisoner has two clocks in his cell, both of which run perfectly accurately. However, clock *A* initially reads 6 PM, clock *B* initially reads 7 PM. The prisoner knows that one of the clocks is set accurately, the other one is one hour off. The prisoner has no idea which one is set accurately; indeed, he initially has degree of belief ½ that *A* is set accurately, and degree of belief ½ that *B* is set accurately. As in the original case, if the coin lands heads the light in his cell will be turned off at midnight, and it will stay on if it lands tails. So initially the prisoner has degree of belief ¼ in each of the following four possible worlds:

W_1: heads and clock *A* is correct
W_2: heads and clock *B* is correct
W_3: tails and clock *A* is correct
W_4: tails and clock *B* is correct.

When in fact it is 11:30 PM the light, for sure, will still be on. What will the prisoner's degrees of belief then be? Well, if the actual world is W_1, then, when it actually is 11:30 PM clock *A* will read 11:30 PM and clock *B* will read 12:30 AM. In that case, since the prisoner sees that the light is still on, he will know that it cannot be that the coin landed heads and clock *B* is correct. That is to say, his degree of belief in W_2 will be 0, and his degrees of belief in the three remaining options will be ⅓ each. Similarly if the actual world is W_3 then at 11:30 PM the prisoner will have degree of belief 0 in W_3 and degree of belief ⅓ in each of the remaining options. On the other hand, if the actual world is W_2 or W_4, then when it is actually 11:30 PM, the clock readings will be 10:30 PM and 11:30 PM, and the prisoner will still have the degrees of belief that he started with, namely ¼ in each of the four possibilities. The prisoner, moreover, knows all of this in advance.

This is rather bizarre, to say the least. For, in the first place, at 6 PM the prisoner knows that at 11:30 PM his degrees of belief in heads will be less or equal to what they now are, and cannot be greater. So his current expectation of what his degrees of belief in heads will be at 11:30 PM is less than his current degree of belief in heads. Second, there is a clear sense in which he does not trust his future degrees

of belief, even though he does not think that he is, or will be, irrational, and even though he can acquire new evidence (the light being on or off). Let D_t denote the prisoner's degrees of belief at time t. Then, for example, $D_{6:00}$(clock B is correct/ $D_{11:30}$ (clock B is correct) = ⅓) = 0. For $D_{11:30}$ (clock B is correct) = ⅓ only occurs in worlds W_1 and W_3, and in each of those worlds clock B is not correct, and the prisoner knows this. Thus his current degrees of belief conditional upon his future degrees of belief do not equal those future degrees of belief. So he systematically distrusts his future degrees of belief. Strange indeed.

IV. SLEEPING BEAUTY

Some researchers are going to put Sleeping Beauty to sleep on Sunday night. During the two days that her sleep will last the researchers will wake her up either once, on Monday morning, or twice, on Monday morning and Tuesday morning. They will toss a fair coin Sunday night in order to determine whether she will be woken up once or twice: if it lands heads she will be woken upon Monday only, if it lands tails she will be woken up on Monday and Tuesday. After each waking, she will be asked what her degree of belief is that the outcome of the coin toss is heads. After she has given her answer she will be given a drug that erases her memory of the waking up; indeed it resets her mental state to the state that it was on Sunday just before she was put to sleep. Then she is put to sleep again. The question now is: When she wakes up, what should her degree of belief be that the outcome was heads?

> *Answer 1*: Her degree of belief in heads should be ½. It was a fair coin and she learned nothing relevant by waking up.
>
> *Answer 2*: Her degree of belief in heads should be ⅓. If this experiment is repeated many times, approximately ⅓ of the awakenings will be heads-awakenings—that is, awakenings that happen on trials in which the coin landed heads.

Adam Elga[3] has argued for the second answer. I agree with him, and I agree with his argument. But let me amplify this view by giving a different argument for the same conclusion. Suppose that Sleeping Beauty is a frequent and rational dreamer. Suppose in fact that every morning if Sleeping Beauty is not woken up at 9 AM, she dreams at 9 AM that she is woken up at 9 AM. Suppose that the dream and reality are indistinguishable in terms of her experience, except that

[3] "Self-Locating Belief and the Sleeping Beauty Problem," *Analysis*, LX (2000): 143–47.

if Sleeping Beauty pinches herself and she is dreaming, it does not hurt (and she does not wake up), while if she does this while she is awake it does hurt. And let us suppose that Sleeping Beauty always remembers to pinch herself a few minutes after she experiences waking up (whether for real, or in a dream.) What should her degrees of belief be when she experiences waking up? It seems obvious she should consider the four possibilities equally likely (the four possibilities being: Monday & Tails & Awake, Monday & Heads & Awake, Tuesday & Tails & Awake, Tuesday & Heads & Dreaming). If Sleeping Beauty then pinches herself and finds herself to be awake, she should conditionalize and then have degree of belief ⅓ in each of the remaining three possibilities (Monday & Tails & Awake, Monday & Heads & Awake, Tuesday & Tails & Awake). Suppose now that at some point in her life Sleeping Beauty loses the habit of dreaming. She no longer needs to pinch herself; directly upon waking she knows that she is not asleep. It seems clear, however, that this lack of dreaming should make no difference as to her degrees of belief upon realizing that she is awake. The process now occurs immediately, without the need for a pinch, but the end result ought to be the same.

Here again, the crucial assumption is commutativity: if the relevant evidence and experience collected is the same, then the order of collection should not matter for the final degrees of belief.[4] But there is clearly something very puzzling about such foreseeable changes in degrees of belief.

V. DUPLICATION

Scenario 1: While you are at the beach, Vishnu tells you that, contrary to appearances, you have existed only for one month: Brahma created you one month ago, complete with all your memories, habits, bad back, and everything. What is more, says Vishnu, one month ago Brahma in fact created two human beings like you (you are one of them), in exactly the same environment, at two different ends of the universe: one on earth, one on twin earth. Unfortunately, Vishnu has a further surprise for you: one month ago Shiva tossed a coin. If it landed heads, Shiva will destroy the human being that is on twin earth one month from now. If it landed tails, Shiva will do nothing. Vishnu does not tell you whether you are to be destroyed, but recommends that if you want to know, you should go check your mail at

[4] Cian Dorr has independently arrived at the idea of using commutativity in order to argue for the degrees of belief that Elga advocates in the Sleeping Beauty case—see Dorr, "Sleeping Beauty: In Defense of Elga," *Analysis* (forthcoming).

home. If there is a letter from President Bush for you, then you will be destroyed. Before running home, what degree of belief should you have in the four possibilities (Earth & Heads, Earth & Tails, Twin Earth & Heads, Twin Earth & Tails)? It seems clear that you should have degree of belief ¼ in each, or at the very least, that it is not irrational to have degree of belief ¼ in each. You run home, and find no letter from Bush. What should your degrees of belief now be? Well, by conditionalization, they should now be ⅓ in each of the remaining possibilities (Earth & Tails, Twin Earth & Heads, Twin Earth & Tails). Consequently you should now have degree of belief ⅓ that the toss landed heads and ⅔ that it landed tails.

Scenario 2: same as scenario 1, except that Vishnu tells you that if the toss came heads, your identical twin was destroyed by Shiva one week ago. Since you were obviously not destroyed, you do not need to rush home to look for a letter from Bush. In essence, you have learned the same as you learned in the previous scenario when you found you had no letter from Bush, and hence you should now have degree of belief ⅓ that the toss landed heads.

Scenario 3: same as scenario 2, except that Vishnu tells you that rather than that two beings were created one month ago by Brahma, one of them already existed and had exactly the life you remember having had. This makes no relevant difference and you should now have degree of belief ⅓ that the coin landed heads.

Scenario 4: same as scenario 3, except that Vishnu tells you that if the coin landed heads one month ago, Shiva immediately prevented Brahma from creating the additional human being one month ago. The upshot is that only if the coin landed tails will Brahma have created the additional human being. Since the timing of the destruction/prevention makes no relevant difference, you should again have degree of belief ⅓ that the coin landed heads.

Scenario 5:[5] you are on earth, and you know it. Vishnu tells you that one month from now Brahma will toss a coin. If it lands tails, Brahma will create, at the other end of the universe, another human being identical to you, in the same state as you will then be, and in an identical environment as you will then be. What do you now think that your degrees of belief should be in one month's time? The answer is that they should be the same as they are in scenario 5, since in one month's time you will be in exactly the epistemic situation that is described in scenario 5. Of course, it is plausible to claim that your

[5] This scenario is similar to the "Dr Evil scenario" in Elga, "Defeating Dr Evil with Self-Locating Belief" (manuscript).

future self will actually be on earth, since it is only your future continuation on earth that can plausibly be called "your future self." That does not mean, however, that your future self can be sure that he is on earth. For your future self will know that he will have the same experiences and memories, whether or not he is on earth or on twin earth, and thus he will not know whether he can trust his memories. Thus you now have degree of belief ½ in heads, and yet you know that in one month's time, you will have degree of belief ⅓. This is bizarre, to say the least.

Yet again, the crucial assumption in this reasoning is commutativity: your final degrees of belief should not depend on the order in which you receive all the relevant experience and evidence. You should end up with the same degree of belief—namely, degree of belief ½ in heads, whether you all along knew you were on earth, or whether you only later found out that you were on earth. But that can only be so if you had degree of belief ⅓ in heads prior to discovering that you were on earth.

VI. DIAGNOSIS

van Fraassen's reflection principle[6] says that one should trust one's future degrees of belief in the sense that one's current degree of belief D_0 in any proposition X, given that one's future degree of belief D_t in X equals p, should be p: $D_0(X/D_t(X)=p)=p$. Given that one is sure that one will have precise degrees of belief at time t, the reflection principle entails that one's current degrees of belief equal the expectations of one's future degrees of belief: $D_0(X)=\Sigma_p D_0(D_t(X)=p)$. The reflection principle is violated in each of the five puzzles that I have presented, for in each case there is a time at which one's expectation of one's future degree of belief in heads differs from one's current degree of belief in heads. This is presumably why we find these cases, prima facie, so worrying and strange.

The source of the problem, I claim, is that the degrees of belief of perfectly rational people, people who are not subject to memory loss or any other cognitive defect, can develop in ways that are as yet unrecognized, and indeed are not allowed according to standard Bayesian lore. Standard Bayesian lore has it that rational people satisfy the principle of conditionalization: rational people alter their degrees of belief only by strict conditionalization on the evidence that they

[6] See van Fraassen, "Belief and the Problem of Ulysses and the Sirens," *Philosophical Studies*, LXXVII (1995): 7–37.

acquire.[7] Strict conditionalization of one's degrees of belief upon proposition X can be pictured in the following manner. One's degrees of belief are a function on the set of possibilities that one entertains. Since this function satisfies the axioms of probability theory it is normalized: it integrates (over all possibilities) to one. Conditionalizing such a function on proposition X then amounts to the following: the function is set to zero over those possibilities that are inconsistent with X, while the remaining nonzero part of the function is boosted (by the same factor) everywhere so that it integrates to one once again. Thus, without being too rigorous about it, it is clear that conditionalization can only serve to "narrow down" one's degree of belief distribution (one really *learns* by conditionalization). In particular a degree of belief distribution that becomes more "spread out" as time passes cannot be developing by conditionalization, and a degree of belief distribution that exactly retains its shape, but is shifted as a whole over the space of possibilities, cannot be developing by conditionalization. Such spreading out and shifting, however, is exactly what occurs in the five puzzles that I presented.

The reasons for such spreading and shifting are very simple. First, let us consider shifting. Suppose that one knows exactly what the history of the world that one inhabits is like. And suppose that one is constantly looking at a clock one knows to be perfect. One's degrees of belief will then be entirely concentrated on one possible world, and at any given moment one's degrees of belief within that world will be entirely concentrated on one temporal location, namely, the one that corresponds to the clock reading that one is then seeing. And that of course means that the location where one's degree of belief distribution is concentrated is constantly moving. That is to say, one's degree of belief distribution is constantly shifting, and such a constant shifting is simply not a case of conditionalization. Self-locating beliefs will therefore generically develop in ways that violate conditionalization. Collins's prisoner case involves exactly such a shifting of one's self-locating degrees of belief. The only difference is that, in his case, one additionally has an initial uncertainty as to which clock is accurate, that is, one is initially uncertain whether one is in a world in which clock A is correct or one in which clock B is correct. It is somewhat surprising that this kind of violation of conditionalization

[7] Strict conditionalization: when one learns proposition X at t, one's new degree of belief D_t equals one's old degree of belief D_0 conditional upon X: $D_t(Y)=D_0(Y/X)$. One might also allow Jeffrey conditionalization. It matters not for our purposes.

can be parlayed into a violation of reflection. But Collins's prisoner case shows exactly how one can do this.

Next, let us consider spreading. The simplest case of spreading is the case of the traveler who takes the path by the Mountains to Shangri La. His degrees of belief become more spread out when he arrives in Shangri La: at that time he goes from degrees of belief one in heads and zero in tails, to degrees of belief ½ in heads and ½ in tails.[8] The reason why this happens is that there are two distinct possible experiential paths that end up in the same experiential state. That is to say, the traveler's experiences earlier on determine whether possibility *A* is the case (Path by the Mountain), or whether possibility *B* is the case (Path by the Ocean). But because of the memory replacement that occurs if possibility *B* is the case, those different experiential paths merge into the same experience, so that that experience is not sufficient to tell which path was taken. Our traveler therefore has an unfortunate loss of information, due to the loss of the discriminating power of his experience. What is somewhat surprising is that this loss of discriminating power is not due to any loss of memory or any cognitive defect on his part: it is due to the fact that something strange would have happened to him had he taken the other path! This loss of discriminatory power of experience, and consequent spreading out of degrees of belief here does not involve self-locating degrees of belief. Suppose, for example, that our traveler is the only person ever to travel along either path. Then our traveler initially is unsure whether he is in a world in which path *A* is never taken or whether he is in a world in which path *B* is never taken. He then becomes sure that he is in a world in which path *B* is never taken. Even later,

[8] van Fraassen, in conversation with me, has suggested that in such situations conditionalization indeed should be violated, but reflection should not. In particular, he suggested that the degrees of belief of the traveler should become *completely vague*, upon arrival in Shangri La. This does not strike me as plausible. Surely upon arrival in Shangri La our traveler is effectively in the same epistemic situation as someone who simply knows that a fair coin has been tossed. One can make this vivid by considering two travelers, *A* and *B*. Traveler *A* never looks out of the window of the car, and hence maintains degree of belief ½ in heads all the way. (The memory replacement device does not operate on travelers who never look out of the window.) Traveler *A*, even by van Fraassen's lights, upon arrival in Shangri La, should still have degree of belief ½ in heads. Traveler *B*, however, does look out of the window during the trip. Upon arrival, by van Fraassen's lights, *B*'s degrees of belief should become completely vague. But it seems odd to me that traveler *B* is epistemically penalized, that is, is forced to acquire completely vague degrees of belief, just because he looked out of the window during the trip, when it seems clear that he ends up in exactly the same epistemic position as his companion, who did not look out of the window.

upon arrival, he again becomes unsure as to which world he is in. None of this has anything to do with self-locating beliefs.[9]

The source of the Sleeping Beauty and Duplication problems is exactly the same. In the case of Sleeping Beauty, the possibility of memory erasure ensures that the self-locating degrees of belief of Sleeping Beauty, even on Monday when she has suffered no memory erasure, become spread out over two days. In the Duplication case, yet again, the possible duplication of experiences forces one to become uncertain as to where (or who) one is. The cause of the spreading of degrees of belief in both cases is "experience duplication," and has nothing to do with the self-locating nature of these beliefs.[10]

It is not very surprising that the spreading of degrees of belief can bring about a violation of reflection. For instance, in the non-self-locating case a predictable reduction from degree of belief one in some proposition X to anything less than one will immediately violate reflection: now you know it, now you do not. The argument is slightly less straightforward in the self-locating case. Consider, for example, a case in which one is on Earth and one knows that at midnight a duplicate of oneself will be created on Mars. One might claim that since one now is certain that one is on Earth, and at midnight one will be uncertain as to whether one is on Earth, thus one has a clear violation of reflection. This is too quick, however. To have a clear violation of reflection it has to be the very same "content of belief" such that one's current degree of belief differs from one's expectation of one's future degree of belief. Depending on what one takes to be the contents of belief when it concerns self-locating beliefs (propositions? maps from locations to propositions?...), one might argue that the

[9] It is obvious how to generalize this case to a case in which there are memory replacement devices at the end of both roads, where these memory replacement devices are indeterministic, that is, when it is the case that for each possible path there are certain objective chances for certain memories upon arrival in Shangri La. For, given such chances (and the principal principle), one can easily calculate the degrees of belief that one should have (in heads and tails) given the memory state that one ends up with. And, generically, one will still violate conditionalization and reflection.

[10] Some people will balk at some of the degrees of belief that I have argued for in this paper, in particular in the self-locating cases. For instance, some people will insist that tomorrow one should still be certain that one is on Earth, even when one now knows (for sure) that a perfect duplicate of oneself will be created on Mars at midnight tonight. I beg to differ. Even if in this case, and other cases, however, one disagrees with me as to which degrees of belief are rationally mandated, the main claim of this paper still stands. The main claim is that in such cases of possible experience duplication, it is at the very least *rationally permissible* that one's degrees of belief become more spread out as time progresses, and hence rational people can violate conditionalization and reflection.

contents of belief are not the same at the two different times, and hence there is no violation of reflection. The arguments of sections IV and V, however, show that one can in any case parlay such spreading of self-locating degrees of belief into violations of reflection concerning such ordinary beliefs as to whether a coin lands heads or tails. So reflection is suckered anyhow.

Finally, the original case of the prisoner involves both a spreading of degrees of belief and a shifting of degrees of belief. The shifting is due simply to the passage of time and the self-locating nature of the beliefs. The spreading is due to the fact that our prisoner does not have experiences that are discriminating enough to pick out a unique location in time.[11] The analysis of section II shows, yet again, that such a spreading and shifting of self-locating degrees of belief can be parlayed into a violation of reflection concerning such ordinary beliefs as to whether a coin lands heads or tails.

VII. CONCLUSIONS

The degrees of belief of rational people can undergo two as yet unrecognized types of development. Such degrees of belief can become more spread out due to the duplication of experiences, or more generally, due to the loss of discriminating power of experiences, and thereby violate conditionalization. In addition, self-locating degrees of belief will generically be shifted over the space of possible locations, due to the passage of time, and thereby violate conditionalization. Such violations of conditionalization can be parlayed into violations of reflection, and lead to a distrust of one's future degrees of belief. Strange, but not irrational.

FRANK ARNTZENIUS

Rutgers University

[11] One might model the prisoner here as having unique distinct experiences at each distinct, external clock, time, and as initially having precise degrees of belief over the possible ways in which those experiences could correlate to the actual, external clock, time. If one were to do so, then the prisoner would merely be initially uncertain as to which world he was in (where worlds are distinguished by how his experiences line up with the actual, external clock, time), but for each such possible world would be always certain as to where he was located in it. And, if one were to do so, then the original prisoner case would be essentially the same case as Collins's prisoner case: no uncertainty of location in any given world, merely an initial uncertainty as to which world one is in, and a subsequent shifting of the locally concentrated degrees of belief within each of the possible worlds. There is no need, however, to represent the original prisoner case that way. Indeed, it seems psychologically somewhat implausible to do so. More importantly, the arguments and conclusions here do not depend on how one models this case.

THE STALNAKER-LEWIS APPROACH TO COUNTERFACTUALS

This general approach to counterfactuals—which appeals to similarity relations between possible worlds—was first set out by Robert Stalnaker in his 1968 article, "A Theory of Conditionals,"[1] and then a modified, and in some ways more satisfactory version of it was advanced by David Lewis in his 1973 book, *Counterfactuals*.[2]

Lewis's detailed exposition, in that book, of this type of account of counterfactuals elicited some very important criticisms of the whole approach, including ones advanced in early reviews by Jonathan Bennett[3] and Kit Fine.[4] But Lewis, in "Counterfactual Dependence and Time's Arrow,"[5] and then in his "Postscripts to 'Counterfactual Dependence and Time's Arrow',"[6] attempted to show that, given, for example, the right sort of account of the factors that enter into judgments of similarity between possible worlds in the case of counterfactuals, one could escape the most crucial objections that had been advanced.

My goal here is to show that, while the answer that Lewis offered to the most fundamental objection advanced by Bennett and Fine is successful in blocking the objection as they stated it, it fails when confronted with a reformulation of that objection.

I. THE STALNAKER-LEWIS ACCOUNTS

In his book, *Counterfactuals*, Lewis offered the following, "ordering relation" account of counterfactuals:

> $\phi \mathbin{\Box\!\!\rightarrow} \psi$ is true at a world i (according to a system of spheres \$) if and only if either
>
> (1) no ϕ-world belongs to any sphere S in \$, or
> (2) some sphere S in \$, does contain at least one ϕ-world, and $\phi \supset \psi$ holds at every world in S (*op. cit.*, p. 16).

What is the "system of spheres"? The basic idea is that for any possible

[1] In Nicholas Rescher, ed., *Studies in Logical Theory: American Philosophical Quarterly*, Monograph 2 (Oxford: Blackwell, 1968), pp. 98–112.

[2] Cambridge: Harvard, 1973.

[3] "Counterfactuals and Possible Worlds," *Canadian Journal of Philosophy*, IV (1974): 381–402.

[4] "Critical Notice: *Counterfactuals*," *Mind*, LXXXIV (1975): 451–58.

[5] *Noûs*, XIII (1979): 455–76.

[6] *Philosophical Papers*, Volume II (New York: Oxford, 1986), pp. 52–66

0022-362X/03/0007/371–77 © 2003 The Journal of Philosophy, Inc.

world, all other possible worlds can be placed on spheres that are centered on the world in question, with the size of a given sphere representing how close each world on the sphere is to the world that lies at the center of the given system of spheres. Thus, all worlds on a given sphere are equally similar to the world at the center, and if one sphere is inside another, then the worlds on the inner sphere are more similar to the world at the center than are the worlds on the outer sphere.

Accordingly, if one replaces the reference to a system of spheres by a direct reference to similarity, the account of the truth conditions of counterfactuals that Lewis is advancing is roughly as follows:

> The counterfactual 'If p were the case, then q would be the case' is true in world W if and only if
>
> Either (1) there is no world at all where p is true, or else (2) some world W^* in which both p and q are true is closer to W than any world in which p is true and q is false.

II. A CRUCIAL OBJECTION TO THE THEORY: BENNETT AND FINE

In his review of Lewis's *Counterfactuals*, Bennett advanced a number of objections to a similarity-over-possible-worlds approach to counterfactuals. But Bennett suggested that the "fatal defect" in the whole approach was that it either generated the wrong truth values for certain counterfactuals, or else it involved unsound judgments of similarity.

Bennett illustrated his point by a counterfactual concerning Lee Harvey Oswald and the death of President Kennedy:

> If Oswald had not killed him, Kennedy would not have been killed.

Suppose, Bennett says, that the Warrenite hypothesis that Oswald acted alone, and so forth, is true. Then Lewis's approach will generate the wrong truth value for the above counterfactual, since it is, Bennett claims, "incredible" that "some worlds in which no one kills Kennedy are more like the actual world than is *any* world in which Kennedy is killed by someone other than Oswald" (*op. cit.*, p. 395).

Similarly, Fine, in his review of *Counterfactuals*, also argued that Lewis's approach generates the wrong truth values for counterfactuals where the consequent could only be true if the world were radically different from the actual world, and Fine illustrated this point by the following counterfactual concerning President Nixon and the button:

> If Nixon had pressed the button there would have been a nuclear holocaust (*op. cit.*, p. 452).

III. LEWIS'S RESPONSE TO THE CRUCIAL OBJECTION

Lewis's response to this objection was set out in his article, "Counterfactual Dependence and Time's Arrow" (*op. cit.*). The thrust of it is that the theory of counterfactuals in question need not be formulated in terms of our ordinary standards of overall similarity. Nor should it be, for then the central objection advanced by Bennett and Fine would be correct. The crucial question, accordingly, is whether one can specify weightings for factors that are relevant to the similarity of one world to another that will combine to produce a measure of overall similarity that will generate the right truth values, and Lewis contends that this is possible.

In particular, Lewis proposes a weighting of factors according to which, while a perfect match of particular facts for an extended stretch of time counts for more with respect to overall similarity than the absence of a single, small, localized miracle, or violation of a law of nature, it is less important than the absence of large miracles. Given this weighting, any world in which Nixon presses the button, if it contains only a small miracle that stops the transmission of the signal, and thus the nuclear holocaust, will not involve a perfect match with respect to future facts—since, for example, the button will be warmer because of contact with Nixon's finger, rays of light in the vicinity of Nixon's hand will be affected differently, Nixon's memories will be different, and so on, all of which will lead to significant divergences with regard to the future state of the world at all later times. One could, of course, consider a world where all of those differences were eliminated by miracles, but then one would be achieving a perfect match at the cost of a large miracle—a cost that, according to Lewis's proposed criteria for overall similarity, is too high.

IV. THE "NIXON AND THE BUTTON" OBJECTION REVISITED

Lewis's response enabled him to escape the specific formulations of the objection in question that were advanced by Bennett and Fine. But, as I shall now show, it cannot handle the more general, underlying objection.

The reason it cannot is that Lewis's proposed solution depends upon the fact that Nixon's pressing the button is an event which has *multiple* effects, and multiple effects of such a sort that it would take a very big miracle to remove all traces of the event, in order to make it the case that there was a perfect match with the future of the actual world. But this means that if it is logically possible to construct a parallel case where the crucial event does not have multiple effects of such a sort that it would take something beyond a small miracle to

remove all traces, Lewis's response will not work, and the fundamental objection will stand.

Can that be done? The answer is that it can be, and in at least two different ways. The first involves considering a world that contains at least one type of causal process that has the following two properties:

(1) The causal process is nonbranching, at least over some temporal interval;
(2) The causal process brings about events that are causally necessary conditions for possible subsequent, branching, causal processes.

Schematically, then, the idea is that one considers a world where there are deterministic causal laws that entail that an event of type C will give rise to a causal process that leads, after a temporal interval t_1, to an event of type E, and where, moreover, at every instant during the relevant temporal interval, there is at most one event that is causally related to the event of type C. Suppose, further, that there is another type of event—D—such that, first, events of type D are causally unrelated to events of type C, and, second, if an event of type E occurs in the presence of an event of type D, the result will be multiply-branching causal processes, leading, after a further temporal interval t_2, to the occurrence of events of types F_1, F_2, F_3,..., F_n. Now let W be a world where such causal laws obtain, but where no event of type C occurs at time t. Assume, however, that an event of type D does occur at time $(t + t_1)$. Consider, then, the following counterfactual, in world W:

(*) If an event of type C had occurred at time t, then events of types F_1, F_2, F_3,..., F_n would have occurred at time $(t + t_1 + t_2)$.

This counterfactual is clearly true in the world we are considering, but it comes out false on Lewis's approach. For consider two worlds, W_1 and W_2, that are otherwise as similar to the original world, W, as possible, but where an event of type C does occur at time t, and where the following propositions are true in the respective worlds:

In W_1: An event of type E occurs at time $(t + t_1)$, and events of types F_1, F_2, F_3,..., F_n occur at time $(t + t_1 + t_2)$

In W_2: No event of type E occurs at time $(t + t_1)$, and no events of types F_1, F_2, F_3,..., F_n occur at time $(t + t_1 + t_2)$.

W_2 differs from W_1 in two respects. First, it involves a single, small violation of a law of the original world, W, since one has an occurrence of an event of type C at time t, but no event of type E at time $(t + t_1)$. In this respect, W_2 is less like the original world than W_1 is. But, secondly, W_2 is a perfect match with W from time $(t + t_1)$ onward,

whereas W_1 diverges from W with the occurrence at time $(t + t_1 + t_2)$ of events of types $F_1, F_2, F_3, \ldots, F_n$, and this divergence will then become ever greater as the resulting causal processes continue to branch. The upshot is that, given the measure of similarity proposed by Lewis, W_2 is closer to the original world, W, than W_1 is, and so it follows, on his account of the truth conditions of counterfactuals, that if an event of type C had occurred at time t, neither an event of type E at time $(t + t_1)$ nor events of types $F_1, F_2, F_3, \ldots, F_n$ at time $(t + t_1 + t_2)$ would have occurred. So counterfactual (*) gets wrongly classified as false.

Formulated in terms of Nixon and the bomb, the example could be as follows: imagine a world that is different from ours in certain respects. First, it is a world where it is possible to bring about physical events psychokinetically. Secondly, it is a world where an act of willing that something be brought about psychokinetically involves no physical change: it consists, instead, only of an appropriate mental state involving emergent qualia. Finally, such a qualia state is almost causally impotent: its only effect is the psychokinetically caused occurrence of the event that was willed; there is not even any memory trace of the relevant act of willing in the person who performed the act.

Here I have formulated things in terms of a direct causal connection between the act of willing and the occurrence of the event willed. If such a direct connection is thought to be somehow unacceptable, one can easily arrange a mechanism: there can be a nonbranching causal chain that proceeds along a straight line to the location where the effect occurs, and where no part of the intervening causal chain has any other effects.

A strange world, no doubt! Yet surely one that is logically possible. But, then, imagine Nixon—or a Nixon counterpart—in a world of this type where he does not will that the button be pressed psychokinetically. What would be the case if Nixon, in such a world, *had* willed that the button be pressed psychokinetically? The correct answer, surely, is given by the following counterfactual:

> If Nixon had willed that the button be pressed psychokinetically, then that would have happened, and there would have been a nuclear holocaust.

But on Lewis's approach, this counterfactual will be false. For an act of willing that something be brought about psychokinetically, in the world that we are considering, will have only one effect: the occurrence of the event that was willed to happen. There is no causal branching, and so only a single, small, localized miracle is required to bring it about that although Nixon has willed that the button be pressed psychokinetically, the button does not move, and thus, rather than

there being a nuclear holocaust, there is, instead, a perfect match with the future of the original world in which Nixon does not will that the button be pressed psychokinetically. So if a single, small, localized miracle contributes less to dissimilarity than a perfect match with respect to all future states of affairs contributes to similarity, then it follows that the above counterfactual is false, rather than true. So Lewis's response fails: it cannot handle a variation on the Nixon and the button example.

A second way of showing that Lewis's response to the type of objection advanced by Bennett and Fine does not work involves appealing, not to the possibility of nonbranching causal processes, but, instead, to the possibility of causal processes that come to an end.

Schematically, the idea is that one considers a world where the occurrence of an event of type C gives rise to causal processes that spread out in every direction, each of which gives rise, after a distance m, and a temporal interval t_1, to an event of type E. On their own, events of type E have no effects whatsoever. An event of type E, however, does have effects when, and only when, it occurs together with an event of type D, and then the result is multiply-branching causal processes, leading, after a further temporal interval t_2, to the occurrence of events of types F_1, F_2, F_3,..., F_n. Now let W be a world where these causal laws obtain, but where no event of type C occurs at location s at time t. Assume, moreover, that there is only a single spatial location, at time $(t + t_1)$, where an event of type D occurs at distance m from location s. Consider, then, the following counterfactual:

(**) If an event of type C had occurred at time t at location s, events of types F_1, F_2, F_3,..., F_n would have occurred at time $(t + t_1 + t_2)$

This counterfactual is clearly true in the world we are considering, for the occurrence of an event of type C would initiate causal processes that spread out in every direction, and, as a consequence, events of type E would occur at time $(t + t_1)$ at various spatial locations, including the one spatial location at that time where an event of type D occurs at distance m from location s. The co-occurrence of an event of type E with an event of type D at the location in question would then result, in turn, in the occurrence of events of types F_1, F_2, F_3,..., F_n at time $(t + t_1 + t_2)$

Counterfactual (**) comes out false, however, on Lewis's approach. For consider two worlds, W_1 and W_2, that are otherwise as similar to the original world, W, as possible, but where an event of type C does occur at time t at location s, and where the following propositions are true in the respective worlds:

In W_1: Events of type E occur at time $(t + t_1)$ at every location at distance m from location s, including the one such location where an event of type D occurs, and events of types F_1, F_2, F_3,..., F_n occur at time $(t + t_1 + t_2)$

In W_2: Events of type E occur at time $(t + t_1)$ at every location at distance m from location s, *except* for the one such location where an event of type D occurs, and no events of types F_1, F_2, F_3,..., F_n occur at time $(t + t_1 + t_2)$

W_2 differs from W_1 in two respects. First, it involves a single, small violation of a law of the original world, W, since one has an occurrence of an event of type C at time t, but no event of type E at time $(t + t_1)$ at the one location at distance m from location s where an event of type D occurs. In this respect, W_2 is less like the original world, W, than W_1 is. But, second, W_2 is a perfect match with W at every moment after time $(t + t_1)$, whereas W_1 diverges from W at every moment from time $(t + t_1 + t_2)$ onward, in view of the occurrence of events of types F_1, F_2, F_3,..., F_n at time $(t + t_1 + t_2)$, and this divergence will then become ever greater as the resulting causal processes continue to branch. The upshot is that, given the measure of similarity proposed by Lewis, W_2 is closer to the original world, W, than W_1 is, and so it follows, on Lewis's account of the truth conditions of counterfactuals, that if an event of type C had occurred at time t, at location s, then there would have been no occurrence either of an event of type E at time $(t + t_1)$ at the one location where an event of type D occurs at distance m from location s, or of any subsequent events of types F_1, F_2, F_3,..., F_n at time $(t + t_1 + t_2)$. So counterfactual (**) gets classified, incorrectly, as false.

The overall conclusion, accordingly, is that Lewis's attempt to answer the most crucial objection to the whole similarity-across-possible-worlds approach to counterfactuals is unsuccessful, since the possibility of worlds that contain either a single nonbranching type of causal process, or else branching causal processes that terminate after a finite time, shows that some counterfactuals get assigned the wrong truth values by a Stalnaker-Lewis account.

MICHAEL TOOLEY

University of Colorado at Boulder

NEW BOOKS: ANTHOLOGIES

ABOULAFIA, MITCHELL, MYRA BOOKMAN, and CATHERINE KEMP, eds.: *Habermas and Pragmatism*. New York: Routledge, 2002. xi, 244 p. Paper $24.95.

ALCOFF, LINDA M., and EDUARDO MENDIETA, eds.: *Identities: Race, Class, Gender, and Nationality*. Malden, MA: Blackwell, 2003. xv, 428 p. Paper $39.95.

ALLMAN, DWIGHT D., and MICHAEL D. BEATY, eds.: *Cultivating Citizens: Soulcraft and Citizenship in Contemporary America*. New York: Lexington, 2002. vii, 167 p. Cloth $70.00, paper $23.95.

ALPERSON, PHILIP, ed.: *Diversity of Community: An Interdisciplinary Reader*. Malden, MA: Blackwell, 2002. xiii, 351 p. Cloth $55.00, paper $16.99.

ALSTON, WILLIAM P., ed.: *Realism and Antirealism*. Ithaca: Cornell, 2002. vi, 303 p. Paper $22.50.

ANTONY, LOUISE M., and NORBERT HORNSTEIN, eds.: *Chomsky and His Critics*. Malden, MA: Blackwell, 2003. viii, 332 p. Paper $29.95.

ARBIB, MICHAEL A., ed.: *The Handbook of Brain Theory and Neural Networks*, Second Edition. Cambridge: MIT, 2003. xvii, 1290 p. Cloth $165.00.

ARIEW, ANDRÉ, ROBERT CUMMINS, and MARK PERLMAN, eds.: *Functions: New Essays in the Philosophy of Psychology and Biology*. New York: Oxford, 2002. viii, 449 p. Cloth $70.00, paper $24.95.

ARRINGTON, ROBERT L., ed.: *The World's Great Philosophers*. Malden, MA: Blackwell, 2003. xiii, 361 p. Paper $21.95.

BAARS, BERNARD J., WILLIAM P. BANKS, and JAMES B. NEWMAN, eds.: *Essential Sources in the Scientific Study of Consciousness*. Cambridge: MIT, 2003. xiii, 1192 p. Paper $60.00.

BACKHAUS, GARY, and JOHN MURUNGI, eds.: *Transformations of Urban and Suburban Landscapes: Perspectives from Philosophy, Geography, and Architecture*. Lanham, MD: Lexington, 2002. vi, 269 p. Cloth $80.00, paper $26.95.

BAGGINI, JULIAN, and JEREMY STANGROOM, eds.: *New British Philosophy: The Interviews*. New York: Routledge, 2002. ix, 303 p. Paper $14.95.

BARHART, MICHAEL, ed.: *Varieties of Ethical Reflection: New Directions for Ethics in a Global Context*. Lanham, MD: Lexington, 2002. vii, 347 p. Cloth $80.00.

BECKMAN, LUDVIG, and EMIL UDDHAMMAR, eds.: *Virtues of Independence and Dependence on Virtues*. New Brunswick, NJ: Transaction, 2003. x, 155 p. Cloth $39.95.

BEILBY, JAMES, ed.: *Naturalism Defeated? Essays on Plantinga's Evolutionary Argument against Naturalism*. Ithaca: Cornell, 2002. x, 283 p. Cloth $45.00, paper $19.95.

BEKOFF, MARC, COLIN ALLEN, and GORDON M. BURGHARDT, eds.: *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition*. Cambridge: MIT, 2002. xxi, 482 p. Paper $38.00.

BERMÚDEZ, JOSÉ LUIS, and ALAN MILLAR, eds.: *Reason and Nature: Essays in the Theory of Rationality*. New York: Oxford, 2002. viii, 286 p. Cloth $124.95.

BLAKE, NIGEL, PAUL SMEYERS, RICHARD SMITH, and PAUL STANDISH, eds.: *The*

Blackwell Guide to the Philosophy of Education. Malden, MA: Blackwell, 2003. xvi, 414 p. Cloth $69.95, paper $34.95.

BODNÁR, ISTVÁN, and WILLIAM W. FORTENBAUGH, eds.: *Eudemus of Rhodes.* New Brunswick, NJ: Transaction, 2002. ix, 383 p. Cloth $69.95.

BORG, EMMA, ed.: *Meaning and Representation.* Malden, MA.: Blackwell, 2002. vi, 117 p. Paper $24.95.

BROADIE, ALEXANDER, ed.: *The Cambridge Companion to the Scottish Enlightenment.* New York: Cambridge, 2003. xvi, 366 p. Cloth $65.00, paper $23.00.

BROOK, ANDREW, and DON ROSS, eds.: *Daniel Dennett.* New York: Cambridge, 2002. xii, 302 p. Cloth $60.00, paper $20.00.

BUNNIN, NICHOLAS, and E.P. TSUI-JAMES, eds.: *The Blackwell Companion to Philosophy,* Second Edition. Malden, MA: Blackwell, 2003. xviii, 951 p. Cloth $69.95, paper $34.95.

BUSS, SARAH, and LEE OVERTON, eds.: *Contours of Agency: Essays on Themes from Harry Frankfurt.* Cambridge: MIT, 2002. xx, 361 p. Cloth $45 00.

CAHN, STEVEN M., ed.: *The Affirmative Action Debate,* Second Edition. New York: Routledge, 2002. xiv, 233 p. Paper $22.95.

CAHN, STEVEN M., and DAVID SHATZ, eds. *Questions about God: Today's Philosophers Ponder the Divine.* New York: Oxford, 2002. x, 174 p. Cloth $17.95.

CAHOONE, LAWRENCE, ed.: *From Modernism to Postmodernism: An Anthology.* Malden, MA: Blackwell, 2003. xv, 622 p. Paper $39.95.

CAMPBELL, NEIL, ed.: *Mental Causation and the Metaphysics of Mind: A Reader.* Orchard Park, NY: Broadview, 2003. 301 p. Paper $26.95.

CAPUTO, JOHN D., ed.: *The Religious.* Malden, MA: Blackwell, 2002. xiii, 322 p. Cloth $64.95.

CARRUTHERS, PETER, STEPHEN STICH, and MICHAEL SIEGAL, eds.: *The Cognitive Basis of Science.* New York: Cambridge, 2002. xii, 409 p. Cloth $70.00, paper $25.00.

CHAN, ALAN K. L., ed.: *Mencius: Contexts and Interpretations.* Honolulu: Hawaii UP, 2002. vii, 328 p. Cloth $55.00.

CLAPIN, HUGH, ed.: *Philosophy of Mental Representation.* New York: Oxford, 2002. xv, 332 p. Cloth $65.00, paper $22.00.

CLARK, GILLIAN, and TESSA RAJAK, eds.: *Philosophy and Power in the Graeco-Roman World: Essays in Honour of Miriam Griffin.* New York: Oxford, 2002. xvii, 348 p. Cloth $72.00.

COHEN, BERNARD, and GEORGE E. SMITH, eds.: *The Cambridge Companion to Newton.* New York: Cambridge, 2002. xiv, 500 p. Cloth $65.00, paper $23.00.

CONANT, JAMES, and URSZULA M. ZEGLEŃ, eds.: *Hilary Putnam: Pragmatism and Realism.* New York: Routledge, 2002. ix, 242 p. Cloth $80.00.

COWAN, JANE K., MARIE-BÉNÉDICTE DEMBOUR, and RICHARD A. WILSON, eds.: *Culture and Rights: Anthropological Perspectives.* New York: Cambridge, 2001. xiv, 258 p. Cloth $65.00, paper $23.00.

CRITCHLEY, SIMON, and ROBERT BERNASCONI, eds.: *The Cambridge Companion to Levinas.* New York: Cambridge, 2002. xxx, 292 p. Cloth $65.00, paper $23.00.

DARLING-SMITH, BARBARA, ed.: *Courage.* Notre Dame: University Press, 2002. xv, 150 p. Cloth $20.00.

DARWALL, STEPHEN, ed.: *Consequentialism.* Malden, MA: Blackwell, 2003. vii, 301 p. Cloth $62.95, paper $27.95; *Contractarianism/Contractualism.* Malden, MA: Blackwell, 2003. vii, 286 p. Cloth $62.95, paper $27.95; *Deontology.* Malden, MA: Blackwell, 2003. vii, 242 p. Cloth $62.95, paper $27.95; *Virtue Ethics.* Malden, MA: Blackwell, 2003. vii, 261 p. Cloth $62.95, paper $27.95.

DOLLING, LISA M., ARTHUR F. GIANELLI, and GLENN N. STATILE, eds.: *The Tests of Time: Reading in the Development of Physical Theory.* Princeton: University Press, 2003. xlii, 716 p. Cloth $99.50, paper $36.95.

DOOLEY, MARK, ed.: *A Passion for the Impossible: John D. Caputo in Focus.* Albany: SUNY, 2003. xxiii, 323 p. Cloth $92.50, paper $29.95.

DOSTAL, ROBERT J., ed.: *The Cambridge Companion to Gadamer.* New York: Cambridge, 2002. xiii, 317 p. Cloth $65.00, paper $23.00.

EILSTEIN, HELENA, ed.: *A Collection of Polish Works on Philosophical Problems of Time and Spacetime.* Norwell, MA: Kluwer, 2002. vii, 148 p. Cloth $55.00.

ELDRIDGE, RICHARD, ed.: *Stanley Cavell.* New York: Cambridge, 2003. x, 248 p. Cloth $60.00, paper $20.00.

FALMAGNE, RACHEL JOFFE, and MARJORIE HASS, eds.: *Representing Reason: Feminist Theory and Formal Logic.* Lanham, MD: Rowman and Littlefield, 2003. viii, 243 p. Cloth $65.00, paper $24.95.

FAULCONER, JAMES E., ed.: *Transcendence in Philosophy and Religion.* Bloomington: Indiana UP, 2003. v, 151 p. Paper $19.95.

FISHKIN, JAMES S., and PETER LASLETT, eds.: *Debating Deliberative Democracy.* Malden, MA: Blackwell, 2003. xiii, 234 p. Cloth $29.95.

FLOYD, JULIET, and SANFORD SHIEH, eds.: *Future Pasts: The Analytic Tradition in Twentieth-Century Philosophy.* New York: Oxford, 2001. xiii, 463 p. Cloth $89.95.

FRANKENBERRY, NANCY K., ed.: *Radical Interpretation in Religion.* New York: Cambridge, 2002. xv, 231 p. Cloth $65.00, paper $24.00.

FREDERICK, ROBERT E., ed.: *A Companion to Business Ethics.* Malden, MA: Blackwell, 2002. xvi, 464 p. Paper $34.95.

FREEMAN, SAMUEL, ed.: *The Cambridge Companion to Rawls.* New York: Cambridge, 2003. xi, 585 p. Cloth $24.00.

FREY, R. G., and CHRISTOPHER HEATH WELLMAN, eds.: *A Companion to Applied Ethics.* Malden, MA: Blackwell, 2003. xviii, 698 p. Cloth $124.95.

GALE, RICHARD M., ed.: *The Blackwell Guide to Metaphysics.* Malden, MA: Blackwell, 2002. vi, 348 p. Cloth $69.95.

GAVIN, WILLIAM J., ed.: *In Dewey's Wake.* Albany: SUNY, 2003. vi, 249 p. Cloth $71.50, paper $23.95.

GEHRING, VERNA V., and WILLIAM A. GALSTON, eds.: *Philosophical Dimensions of Public Policy.* New Brunswick, NJ: Transaction, 2002. xii, 349 p. Cloth $89.95.

GENDLER, TAMAR SZABÓ, and JOHN HAWTHORNE, eds.: *Conceivability and Possibility.* New York: Oxford, 2002. viii, 507 p. Cloth $72.00, paper $29.95.

GOLOMB, JACOB, and ROBERT S. WISTRICH, eds.: *Nietzsche: Godfather of Fascism?* Princeton: University Press, 2002. xv, 341 p. Cloth $65.00, paper $16.95.

GOODCHILD, PHILIP, ed.: *Rethinking Philosophy of Religion: Approaches from Continental Philosophy.* New York: Fordham, 2002. vi, 339 p. Cloth $40.00.

GOULD, CAROL C., ed.: *Constructivism and Practice: Toward a Historical Epistemology.* Lanham, MD: Rowman and Littlefield, 2003. xii, 272 p. Cloth $75.00, paper $28.95.

GRACIA, JORGE J. E., CAROLYN KORSMEYER, and RODOLPHE GASCHÉ, eds.: *Literary Philosophers: Borges, Calvino, Eco.* New York: Routledge, 2002. vii, 248 p. Paper $23.95.

GRACIA, JORGE J. E., and TIMOTHY B. NOONE, eds.: *A Companion to Philosophy in the Middle Ages.* Malden, MA: Blackwell, 2003. xxi, 739 p. Cloth $124.95.

GRACIA, JORGE J. E., GREGORY M. REICHBERG, and BERNARD N. SCHUMACHER, eds.: *The Classics of Western Philosophy: A Reader's Guide.* Malden, MA: Blackwell, 2003. xviii, 614 p. Paper $29.95.

GREIFF, PABLO DE, and CIARAN CRONIN, eds.: *Global Justice and Transnational Politics: Essays on the Moral and Political Challenges of Globalization.* Cambridge: MIT, 2002. vi, 320 p. Paper $22.95.

GROENHOUT, RUTH E., and MARYA BOWER, eds.: *Philosophy, Feminism, and Faith.* Bloomington: Indiana UP, 2003. vi, 302 p. Cloth $49.95, paper $24.95.

GROSSMAN, JOAN DELANEY, and RUTH RISCHIN, eds.: *William James in Russian Culture.* Lanham, MD: Lexington, 2003. xi, 259 p. Cloth $75.00, paper $28.95.

GUNTHER, YORK H., ed.: *Essays on Nonconceptual Content.* Cambridge: MIT, 2003. x, 339 p. Paper $30.00.

GUTTENPLAN, SAMUEL, JENNIFER HORNSBY, and CHRISTOPHER JANAWAY, eds.: *Reading Philosophy: Selected Texts with a Method for Beginners.* Malden, MA: Blackwell, 2003. xi, 300 p. Paper $27.95.

HAAKONSSEN, KNUD, ed.: *Adam Smith: The Theory of Sentiments.* New York: Cambridge, 2002. xxxi, 411 p. Cloth $65.00, paper $23.00.

HECHT, HEIKO, ROBERT SCHWARTZ, and MARGARET ATHERTON, eds.: *Looking into Pictures: An Interdisciplinary Approach to Pictorial Space.* Cambridge: MIT, 2003. xviii, 417 p. Cloth $55.00.

HAMILTON, CAROLYN, ed.: *Refiguring the Archive.* Norwell, MA: Kluwer, 2002. 368 p. Cloth $98.00.

HARRIS, LEONARD, SCOTT L. PRATT, and ANNE S. WATERS, eds.: *American Philosophies: An Anthology.* Malden, MA: Blackwell, 2002. viii, 456 p. Paper $34.95.

HARRISON, ROSS, ed.: *Henry Sidgwick.* New York: Oxford, 2001. v, 122 p. Cloth $29.95.

HEYES, CRESSIDA J., ed.: *The Grammar of Politics: Wittgenstein and Political Philosophy.* Ithaca, NY: Cornell, 2003. ix, 259 p. Cloth $45.00, paper $19.95.

HILL, DARRYL B., and MICHAEL J. KRAL, eds.: *About Psychology: Essays at the Crossroads of History, Theory, and Philosophy.* Albany: SUNY, 2003. vii, 182 p. Cloth $54.50, paper $17 95.

HILL, THOMAS E., ed.: *Human Welfare and Moral Worth.* New York: Oxford, 2002. vii, 415 p. Cloth $60.00, paper $19.95.

HOWARD-SNYDER, DANIEL, and PAUL K. MOSER, eds.: *Divine Hiddenness: New Essays.* New York: Cambridge, 2002. 242 p. Cloth $60.00, paper $22.00.

HOWIE, JOHN, ed.: *Ethical Issues for a New Millennium.* Carbondale: Southern Illinois UP, 2002. xvi, 219 p. Cloth $35.00.

HUEMER, MICHAEL, ed.: *Epistemology. Contemporary Readings.* New York: Routledge, 2002. xi, 619 p. Paper $27.95.

IRWIN, WILLIAM, ed.: *The Matrix and Philosophy: Welcome to the Desert of the Real.* Chicago: Open Court, 2002. 288 p. Cloth $24.95, paper $17.95.

JACOBS, BRIAN, and PATRICK KAIN, eds.: *Essays on Kant's Anthropology.* New York: Cambridge, 2003. ix, 265 p. Cloth $60.00.

KATZ, RUTH, and RUTH HACOHEN, ed.: *The Arts in Mind: Pioneering Texts of a Coterie of British Men of Letters.* New Brunswick, NJ: Transaction, 2003. vi, 432 p. Cloth $59.95.

KELLER, CATHERINE, and ANNE DANIELL, eds.: *Process and Difference: Between Cosmological and Poststructuralist Postmodernisms.* Albany: SUNY, 2002. xxiii, 272 p. Cloth $62.50, paper $20.95.

KITCHING, GAVIN, and NIGEL PLEASANTS, eds.: *Marx and Wittgenstein: Knowledge, Morality, and Politics.* New York: Routledge, 2002. xiv, 303 p. Cloth $95.00.

KRAUSZ, MICHAEL, ed.: *Is There a Single Right Interpretation?* University Park: Penn State UP, 2002. vi, 423 p. Cloth $85.00, paper $29.50.

LAFOLLETTE, HUGH, ed.: *The Oxford Handbook of Practical Ethics.* New York: Oxford, 2003. xvii, 772 p. Cloth $99.00.

LAGERLUND, HENRIK, and MIKKO YRJÖNSUURI, eds.: *Emotions and Choice from Boethius to Descartes.* Boston: Kluwer, 2002. ix, 342 p. Paper $38.00.

LANDESMAN, CHARLES, and ROBIN MEEKS, eds.: *Philosophical Skepticism.* Malden, MA: Blackwell, 2003. xi, 364 p Paper $27.95.

LEVER, TREVOR, and GERARD L'E TURNER, eds.: *Discussing Chemistry and Steam: The Minutes of a Coffee House Philosophical Society, 1780–1787.* New York: Oxford, 2002. viii, 284. Cloth $110.00.

LEVITIN, DANIEL J., ed.: *Foundations of Cognitive Psychology: Core Readings.* Cambridge: MIT, 2002. xvi, 862 p. Paper $50.00.

LIECHTY, DANIEL, ed.: *Death and Denial: Interdisciplinary Perspectives on the Legacy of Ernest Becker.* Westport, CT: Praeger, 2002. xvi, 301 p. Cloth $66.95.

LIGHT, ANDREW, and AVNER DE-SHALIT, eds.: *Moral and Political Reasoning in Environmental Practice.* Cambridge: MIT, 2003. viii, 257 p. Paper $27.95.

LIGHT, ANDREW, and HOLMES ROLSTON, eds.: *Environmental Ethics: An Anthology.* Malden, MA: Blackwell, 2003. x, 554 p. Cloth $74.95.

LIU, XIUSHENG, and PHILIP J. IVANHOE, eds.: *Essays on the Moral Philosophy of Mengzi.* Indianapolis: Hackett, 2002. xiv, 249 p. Cloth $39.95, paper $16.95.

LOTT TOMMY L., and JOHN P. PITTMAN, eds.: *A Companion to African-American Philosophy.* Malden, MA: Blackwell, 2003. xv, 465 p. Cloth $124.95.

LOUX, MICHAEL J., ed.: *Metaphysics: Contemporary Readings.* New York: Routledge, 2001. xii, 555 p. Cloth $90.00, paper $25.95.

LOVINK, GEERT, ed.: *Uncanny Networks: Dialogues with the Virtual Intelligentsia.* Cambridge: MIT, 2002. xv, 374 p. Cloth $27.95.

MALAMENT, DAVID B., ed.: *Reading Natural Philosophy: Essays in the History and Philosophy of Science and Mathematics.* Chicago: Open Court, 2002. xi, 424 p. Cloth $42.95, paper $32.95.

Derrida and Habermas

New from Chicago

Philosophy in a Time of Terror

Dialogues with Jürgen Habermas and Jacques Derrida

Giovanna Borradori

"This is in many ways the first real engagement between Derrida and Habermas. . . . Each has some searchingly original and important things to say about the terrorist attacks and their political aftermath. . . . Borradori has done an excellent job in editing the interviews and drawing her collocutors out on a range of issues that neither has previously discussed with so powerful a sense of moral and political urgency."—Christopher Norris, Cardiff University

Cloth $25.00

The Problem of Genesis in Husserl's Philosophy

Jacques Derrida

Translated by Marian Hobson

Derrida's first book-length work, *The Problem of Genesis in Husserl's Philosophy*, was originally written for his *diplome d'etudes superieures*. Derrida surveys Husserl's major works on phenomenology, revealing what he sees as an internal tension in Husserl's central notion of "genesis," and gives us our first glimpse into the concerns and frustrations that would later lead Derrida to abandon phenomenology and develop his now famous method of deconstruction.

Cloth $27.50

The University of Chicago Press

1427 East 60th Street, Chicago, Illinois 60637 www.press.uchicago.edu

THE JOURNAL OF PHILOSOPHY

The Journal is pleased to announce to our readers and subscribers the launch of its new website, which will have the pages listed below. The address is: *www.journalofphilosophy.org* or simply *www.jphil.org*.

~•~

PURPOSE & HISTORY
CONTACTS AND ADDRESSES
BOARD OF TRUSTEES
EDITORIAL BOARD
MEMOS TO AUTHORS
SUBSCRIPTIONS & ADVERTISING
RECENT & FORTHCOMING ISSUES
ELECTRONIC ISSUES
INDEX 1904–PRESENT
RIGHTS AND PERMISSIONS
LECTURES & CONFERENCES
PHILOSOPHY LINKS

~•~

Philosophy New from Chicago

Theory and Reality

An Introduction to the Philosophy of Science

Peter Godfrey-Smith

Intended for undergraduates and general readers with no prior background in philosophy, *Theory and Reality* combines a survey of recent history of the philosophy of science with current key debates in language that any beginning scholar or critical reader can follow.

"A well-organized and excellent text. Godfrey-Smith's writing is admirably clear and he deals with very complex philosophical issues with little jargon."—Ron Amundson, University of Hawaii

Paper $24.00

Metaromanticism

Aesthetics, Literature, Theory

Paul Hamilton

"Hamilton offers a searching commentary on key Romantic thinkers to locate them historically and map out the relevance of Romantic thought for contemporary debates about aesthetics. . . . This book is essential reading for anyone who thought that Romantic aesthetics had played out its political hand."—Frances Ferguson, Johns Hopkins University

Paper $25.00

German Idealism and the Jew

The Inner Anti-Semitism of Philosophy and German Jewish Responses

Michael Mack

"A major addition to the critical literature of German Jewry from the Enlightenment to the Shoah. Mack's extraordinary study of the philosophical underpinnings of German anti-Semitism from Kant to Hegel picks up a thread of intellectual corruption that formed major Jewish thinkers from Rosenzweig and Benjamin and beyond. . . . Mack offers a sense of the productive deformation of German Jewish thought through its struggle with German philosophy."—Sander L. Gilman, University of Illinois at Chicago

Cloth $35.00

The End of Philosophy

Martin Heidegger

Translated by Joan Stambaugh

"These essays give the best available picture of Heidegger's doctrine of the history of Being, and are thus helpful towards understanding his thought particularly in its later phases. . . . The last, 'Overcoming Metaphysics,' is especially good on the point of the place of modern technology in the history of Being."—*Library Journal*

Paper $13.00

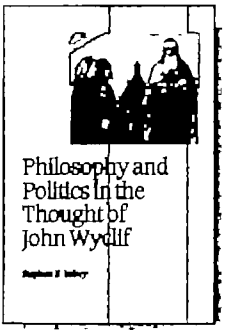


The University of Chicago Press

1427 East 60th Street, Chicago, Illinois 60637 www.press.uchicago.edu

OUTSTANDING SCHOLARSHIP FROM CAMBRIDGE

Aquinas: Political Writings
Edited by R.W. Dyson
Cambridge Texts in the History of Political Thought
0-521-37595-9 Paperback $22.00



Philosophy and Politics in the Thought of John Wyclif
Stephen E. Lahey
Cambridge Studies in Medieval Life and Thought
Fourth Series 54
0-521-63346-X Hardback $60.00

Donald Davidson
Edited by Kirk Ludwig
Contemporary Philosophy in Focus
0-521-79043-3 Hardback $60.00
0-521-79382-3 Paperback $20.00



Knowledge and Indifference in English Romantic Prose
Tim Milnes
Cambridge Studies in Romanticism 55
0-521-81098-1 Hardback $60.00

Alasdair MacIntyre
Edited by Mark C. Murphy
Contemporary Philosophy in Focus
0-521-79042-5 Hardback $65.00
0-521-79381-5 Paperback $20.00



Logic, Thought and Language
Edited by Anthony O'Hear
Royal Institute of Philosophy Supplements 51
0-521-52966-2 Paperback $20.00

A Question of Trust
The BBC Reith Lectures 2002
Onora O'Neill
0-521-82304-8 Hardback $40.00
0-521-52996-4 Paperback $15.00

Visions of Politics
3-Volume Set
Quentin Skinner
0-521-89075-6 Paperback $65.00

Available in bookstores or from

CAMBRIDGE UNIVERSITY PRESS 800-872-7423 us.cambridge.org/philosophy

new books from the mit press

Transcritique

On Kant and Marx
Kojin Karatani
translated by Sabu Kohso
"An immensely ambitious theoretical edifice in which new relations between Kant and Marx are established, as well as a new kind of synthesis between Marxism and anarchism." — Fredric Jameson, Duke University
336 pp $35

Pragmatic Bioethics

second edition
edited by Glenn McGee
Applies the American philosophical theory of pragmatism to ethical issues in health care and biomedical research. The essays present different approaches to pragmatism and different ways of applying pragmatism to scientific and medical matters
Basic Bioethics series • A Bradford Book
320 pp $26.95 paper

Imagination and the Meaningful Brain

Arnold H. Modell
"Modell argues convincingly that subjective experience provides an invaluable window onto the workings of the human brain. He also demonstrates the unexpected usefulness of psychoanalytical theories and methods in tackling some knotted problems of contemporary neuroscience." — Mark Solms, University College, London
A Bradford Book • 272 pp, 5 illus $22

New Essays on Semantic Externalism and Self-Knowledge

edited by Susana Nuccetelli
Essays on the consequences of semantic externalism for knowledge of mind and the empirical world and for our understanding of transmission of epistemic warrant by inference.
A Bradford Book • 280 pp $35

Imagination and Its Pathologies

edited by James Phillips and James Morley
"*Imagination and Its Pathologies* is a veritable feast of philosophical, clinical, historical and cross-cultural scholarship. It is a truly seminal contribution to the rapidly expanding cross-disciplinary canon of philosophy and psychiatry."
— KWM Fulford, University of Warwick
Philosophical Psychopathology Disorders in Mind series
368 pp, 8 illus. $45

Evolution and Learning

The Baldwin Effect Reconsidered
edited by Bruce H. Weber and David J. Depew
Essays on the contributions to historical and contemporary evolutionary theory of the Baldwin effect, which postulates the effects of learned behaviors on evolutionary change.
Life and Mind series • A Bradford Book
352 pp, 12 illus $45

Nature by Design

People, Natural Process, and Ecological Restoration
Eric Higgs
"A wonderful book—an eloquent, wise, and useful guide to the potentials and ambiguities of ecological restoration."
— David W. Orr, Oberlin College
416 pp, 18 illus. $27.95 paper

zone books

Action and Reaction

The Life and Adventures of a Couple
Jean Starobinski
translated by Sophie Hawkes
A study of the word pair "action and reaction" embracing philosophy, semantics, literature, and the history of science
Distributed for Zone Books
450 pp $34

http://mitpress.mit.edu

Paperback

Sublime Understanding

Aesthetic Reflection in Kant and Hegel
Kirk Pillow

Drawing an aesthetics of sublimity from Kant's *Critique of Judgment*, and bolstering it with help from Hegel, Kirk Pillow makes sublimity the center of an alternative conception of aesthetic response and interpretation

Studies in Contemporary German Social Thought series • 385 pp $22 paper

Communicative Action and Rational Choice

Joseph Heath

"An original and important contribution to the literature on Habermas and the theory of communicative action."
— James Bohman, Saint Louis University

Studies in Contemporary German Social Thought series • 375 pp , 12 illus $20 paper

When Self-Consciousness Breaks

Alien Voices and Inserted Thoughts
G. Lynn Stephens and George Graham

"An admirable contribution to the psychological and philosophical literature, one that demonstrates the ways in which philosophy can inform the interdisciplinary study of consciousness." — *Choice*

Philosophical Psychopathology series
A Bradford Book • 212 pp. $16.95 paper

Cycles of Contingency

Developmental Systems and Evolution
edited by Susan Oyama, Paul E. Griffiths, and Russell D. Gray

Provides historical background to developmental systems theory, which offers a way to integrate an evolutionary perspective into the biological and social sciences while avoiding traditional oppositions.

Life and Mind Philosophical Issues in Biology and Psychology series
A Bradford Book • 391 pp , 26 illus. $27 paper

Knowledge, Possibility, and Consciousness

John Perry

"Anyone interested in the philosophy of mind will want to read this book."
— *Choice*

Jean Nicod series • 237 pp $15 paper

Rationality in Action

John R. Searle

"A significant contribution to the philosophy of rational action, and its implications for the problem of free will."
— David Hodgson, *Journal of Consciousness Studies*

Jean Nicod series • 319 pp. $16 95 paper

Norms of Nature

Naturalism and the Nature of Functions
Paul Sheldon Davies

"Davies does an admirable job of explaining and defending his particular vision of [the systemic capacity] theory and its place in the natural sciences." — *Choice*

A Bradford Book • 250 pp, 7 illus $17.95 paper

The Love of Nature and the End of the World

The Unspoken Dimensions of Environmental Concern
Sherry Weber Nicholsen

"Sherry Nicholsen has fashioned a warm meditation on our chilling capacity for oblivion." — David Abram, author of *The Spell of the Sensuous*

226 pp $16.95 paper

Bringing the Biosphere Home

Learning to Perceive Global Environmental Change
Mitchell Thomashow

"This exceedingly engaging and readable book effectively weaves personal narrative and commentary on scientific and literary works." — *Choice*

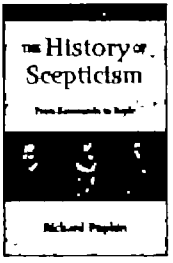
256 pp $15 95 paper

To order call 800-405-1619
Prices subject to change without notice

Now from
Oxford







Prices are subject to change and apply only in the US To order, or for more information, please call 1-800-451-7556
Visit our website at www.oup-usa.org

OXFORD

New in Paperback
Picturing the Human
The Moral Thought of Iris Murdoch
MARIA ANTONACCIO
"Offers a systematic framework for assessing Iris Murdoch's remarkable philosophy...What emerges is the depth and insight of Murdoch's moral imagination. Highly recommended." —*Library Journal*
paper 2003 $22 00

New in Paperback
The Ethics of Killing
Problems at the Margins of Life
JEFF MCMAHAN
"*The Ethics of Killing* is detailed, careful, comprehensive, and innovative [It] is an example of philosophy at the highest level. It is a genuine pleasure to have the opportunity to read such a probing, careful, analytical, honest, and utterly wonderful book."—*Ethics*
paper 2003 $24 95

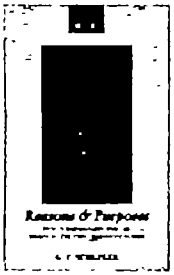
The Practice of Value
JOSEPH RAZ
Edited by R. JAY WALLACE
This volume inaugurates a new interdisciplinary series of books based on the Tanner Lectures given at the University of California, Berkeley. The series aims to make these distinguished lectures, and the lively debates stimulated by their presentation in Berkeley, available to a broad readership
2003 $19 95

The History of Scepticism
From Savonarola to Bayle
Revised Edition
RICHARD H. POPKIN
In this updated version of the classic first published in 1960, Popkin has revised and expanded throughout, and has added three new chapters, one on Savonarola, one on Henry More and Ralph Cudworth, and one on Pascal
2003 paper $24 95

New in Paperback
Logical Properties
Identity, Existence, Predication, Necessity, Truth
COLIN MCGINN
Logicians and philosophers have tried to define such concepts as existence, identity, and truth according to preconceived ideas. McGinn respects the ordinary ways we talk and think when we employ these concepts, while showing that they are far more interesting and peculiar than some have assumed.
paper 2003 $19 95

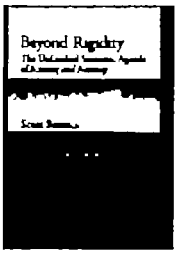
Motivation and Agency
ALFRED R. MELE
What place does motivation have in the lives of intelligent agents? Mele's answer is sensitive to the concerns of philosophers of mind and moral philosophers and informed by empirical work. He offers a distinctive, comprehensive, attractive view of human agency
2003 $49 95

New from
Oxford



Reasons and Purposes
Human Rationality and the
Teleological Explanation of Action
FRED SCHUELER
In this text, Schueler rejects the dominant view that the beliefs and desires that constitute our reasons for acting simply *cause* us to act as we do, and argues instead for a view centered on practical deliberation—our ability to evaluate the reasons we accept.
2003 $39 95

Rules, Reasons, and Norms
Selected Essays
PHILIP PETTIT
Pettit presents a selection of essays touching upon metaphysics, philosophical psychology, and the theory of rational regulation. He discusses the rule-following character of thought, how choice can be responsive to different factors—while still being under the control of thought—and goes on to examine the implications of this view of choice and rationality for the normative regulation of social behavior.
2003 paper $24 95



New in Paperback!
Beyond Rigidity
The Unfinished Semantic Agenda
of Naming and Necessity
SCOTT SOAMES
"This excellent book is aptly titled, for in it Scott Soames systematically discusses and greatly extends the semantic views that Saul Kripke presented in *Naming and Necessity*. As Soames does this, he touches on a wide variety of semantic topics, all of which he treats with his characteristically high degree of clarity, depth, and precision."—David Braun, *Linguistics and Philosophy*
paper 2003 $27 50



Living with Nietzsche
What the Great "Immoralist" Has to Teach Us
ROBERT C. SOLOMON
Rather than focus on the "true" Nietzsche or trying to determine "what Nietzsche really meant" by his seemingly random and often contradictory pronouncements about "the Big Questions" of philosophy, Solomon reminds us that Nietzsche is not a philosopher of abstract ideas but rather of the dazzling personal insight, the provocative challenge, the incisive personal probe. Ultimately, Solomon argues, Nietzsche is an example as well as a promulgator of "passionate inwardness," a life distinguished by its rich passions, exquisite taste, and a sense of personal elegance and excellence.
2003 $35 00

War and Self-Defense
DAVID RODIN
When is it right to go to war? The most persuasive answer to this question has always been 'in self-defense' In a penetrating new analysis, bringing together moral philosophy, political science, and law, David Rodin shows what's wrong with this answer.
2003 $35 00

Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556
Visit our website at www.oup-usa.org

OXFORD

ersity Library
LIFORNIA
MPUS

ADORNO ANSCOMBE AYER BACHELARD
BERGSON BLANSHARD BRAITHWAITE
CASSIRER CAVELL CHISHOLM CHOMSKY
CHURCH CHURCHLAND DAVIDSON
DENNETT DERRIDA DEWEY DUCASSE
DUMMETT DWORKIN EVANS
FEYERABEND FODOR FOOT FRANKENA
FRANKFURT GEACH GETTIER GOLDMAN
GOODMAN HABERMAS HACKING HEMPEL
HINTIKKA JAMESON KAPLAN KATZ KIM
KRIPKE KUHN LANGER LEWIS
MACINTYRE MALCOM MARCUS MARCUSE
MCDOWELL MEAD NAGEL NUSSBAUM
NOZICK PARFIT PEIRCE PLANTINGA
PUTNAM QUINE RAWLS REICHENBACH
RICOEUR RORTY ROYCE RUSSELL
SANTAYANA SARTRE SCANLON SEARLE
SELLARS SEN SHARP STRAWSON
SWINBURNE TAYLOR THOMSON VAN
FRAASSEN VEATCH WILLLIAMS
WITTGENSTEIN WOLLHEIM

THE JOURNAL OF PHILOSOPHY

Celebrates 100 Years

Special Limited Time Offer

Full Volumes 1904-1999

$50.00 per Volume

All volumes are available for $50.00 plus $5.00 shipping and handling.

Send check or money order to:

The Journal of Philosophy
1150 Amsterdam Avenue, MC 4952
Columbia University
New York, NY 10027
(212) 666-4113
www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY



VOLUME C, NUMBER 8
AUGUST 2003

page

383 *On the Foundations of Pragmatic Arguments* — John Cantwell
403 *From Consequentialism to Utilitarianism* — Martin Peterson

COMMENTS AND CRITICISM

416 *The First Person: Error through Misidentification, the Split between Speaker's and Semantic Reference, and the Real Guarantee* — Annalisa Coliva

432 NEW BOOKS: ANTHOLOGIES
435 NEW BOOKS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E WOODBRIDGE AND WENDELL T BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX. (212) 932-3721.

You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2003

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset and printed by Capital City Press, Montpelier, VT.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying More information on our terms may be obtained by consulting our January issue or by writing to us

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

POSTMASTER: Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2003 by the Journal of Philosophy, Inc

ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY
VOLUME C, NO. 8, AUGUST 2003

ON THE FOUNDATIONS OF PRAGMATIC ARGUMENTS*

Of course, one can judge others by any set of standards one chooses, but is there a set of standards that will guarantee that the judgment has normative force? According to a fairly influential tradition there is such a set: the set of standards of the very subject one is judging. In this tradition the question is not whether particular beliefs or preferences are bad, but whether whole systems of beliefs and preferences *cohere*. Why should coherence of beliefs and preferences matter? One answer is that an agent who acts on incoherent beliefs and preferences will do badly and do so by his very own standards. An argument that tries to show this is called a *pragmatic argument*. It is termed *pragmatic* because it appeals to practical consequences of a system of beliefs and preferences rather than to intuitions about whether that system violates some structural constraint that we consider intuitively reasonable.

I

I wish to defend two theses:

(1) The preferences of the agent are *incoherent* if and only if there are situations where the agent judges it bad (or at any rate has reasons to judge it bad) to follow his own preferences.
(2) A pragmatic argument is *sound*—has normative force—if and only if it establishes that the preferences of the agent are incoherent.

The first thesis is perhaps better seen as a stipulative and somewhat vague definition; 'coherence' is a term with many uses and it would be pointless to try to argue that the way I will use it is the only correct one. In sections II, III, and IV, I try to make a reasonably rigorous

* I would like to thank Wlodek Rabinowicz and Isaac Levi for several valuable comments.

0022-362X/03/0008/383–402 © 2003 The Journal of Philosophy, Inc.

formal explication of the definition within the framework of *deterministic* decision situations (situations where there are no chance events).

The second thesis has more substance and it will take some argument to show that it is reasonable. *Dutch books* and *money pumps* are perhaps the best known forms of pragmatic arguments. In the way they are usually presented they involve different kinds of *exploitability*: an agent with incoherent preferences can be made to suffer a sure monetary loss (and this is what makes the preferences incoherent). I will argue that establishing exploitability is neither a necessary nor sufficient condition for establishing the soundness of pragmatic arguments. Preferences that there are good reasons to judge rational can, under certain conditions, be subjected to a pragmatic argument showing that they can be exploited. Furthermore, standard ways of blocking these arguments block too much.

I will argue that exploitability is a sign of irrationality only if it is due to incoherence. As incoherence does not entail exploitability, this may seem to remove that very aspect of incoherence that would establish its normative force: pragmatic considerations. But I will argue that the tie between incoherence and the normatively relevant aspects of pragmatic considerations, as properly understood, remains intact. The tie to be severed is instead that between exploitability and the normatively relevant aspects of pragmatic considerations. This question is closely related to that of *feasibility* and will be discussed in sections VI and VII.

On the surface, pragmatic arguments are directed against particular combinations of beliefs and preferences but they are perhaps better thought of as directed against decision methods that form preferences when there are various dimensions of preference that need to be amalgamated.[1] For instance, there are methods for amalgamating the many dimensions of value associated with an outcome of an act into a single measure to enable comparison with other acts. There are also theories that describe how preferences on the many possible outcomes of acts should shape our preferences on the acts themselves. Pragmatic arguments can establish that such decision methods can give rise to incoherent preferences, but of course one can also use

[1] A more neutral description might be that the decision methods in question provide a structured means for guiding choice in the face of more basic *preferences*. Some hold that there is no such distinction: in choice one's preferences are revealed. I will not review the virtues and drawbacks with either approach, mainly because I think they are irrelevant to the problems I wish to discuss. I will employ the terminology of "revealed preference," but the conclusions drawn do not depend on this.

pragmatic arguments to establish the soundness of the decision methods by showing that they cannot give rise to incoherent preferences.

II

Let S denote a set of *states*. A state u determines all the relevant features of the world with respect to some decision situation: the actions available, the values at stake, and so forth. An *action* takes one from one state to another. The actions available at a state u are given by the set of states $A(u)$, that is, each element v of $A(u)$ can be thought of as an action available at u: the action that brings one to the state v. A state where no action is available ($A(u) = \varnothing$) is called an *end state* or a *final outcome*.

A (*possible*) *action sequence* is an ordered n-tuple (possibly empty) of states $<u_1,\ldots, u_n>$ such that for each $i < n$: $u_{i+1} \in A(u_i)$. I am assuming that every action sequence is finite. A state v can be *reached* from u, if there is some action sequence starting with u and ending with v. When p is the action sequence $<u_1,\ldots, u_n>$, let $p(i)$ denote the ith element of p (that is, u_i) and let $p(\#)$ denote the final element of p.

A *plan* is any action sequence that ends with an end state. Let P_u denote the set of plans available at u, that is, a plan p is an element of P_u if and only if the first element of p is an element of $A(u)$ (so u will not be an element of p, but some element of $A(u)$ will be the first element of p). Let $P_u(\#)$ denote the set $\{p(\#) \mid p \in P_u\}$.

Each state u determines the preferences of the agent. The preferences of an agent at u will be represented by a choice function C_u that has at its domain (i) any finite set of states, (ii) any finite set of plans, and (iii) any finite set of actions (when it is immaterial which particular state u is being referred to I will drop the subscript). There is thus a distinction between preferences on states, preferences on plans, and preferences on actions. I will assume, however, that the end state of a plan is the *prime conveyor of value* for the plan. By this rather cryptic phrase I wish to exclude a number of different possibilities: that the intermediary states in a plan are relevant for the evaluation of a plan; that end states are valued by the plan to which they belong rather than vice versa; or that there are, say, moral reasons for choosing a plan that go over and beyond its consequences.[2] This is a drastic move as it is obvious, for instance, that in decision situations

[2] In particular, it excludes the possibility of the agent preferring to be resolute: having the preference that he always will execute any plan that has been previously adopted. Such a preference involves other values than preferences for outcomes as prime conveyors of value—see E. F. McClennen, *Rationality and Dynamic Choice* (New York: Cambridge, 1990)

that stretch over some time, we are not only interested in the end result but also in how much we enjoy the journey to the end result.

I will assume that $C(X)$ is nonempty whenever X is nonempty. I will also assume the following closure condition: if X is a finite set of states and C is a choice function, then there is a state u such that $A(u) = X$ and $C_u = C$. This guarantees that for any set of preferences there is a potential decision situation where the preferences could be put to a test. This assumption is unnecessarily strong and could with only technical inconvenience be considerably weakened.

Note in particular: (1) the preferences of the agent need not remain constant; (2) as the preferences are relativized to a state, the agent will at u have preferences about what it would be best to do given the alternatives available at some other state v and these need not coincide with the agent's preferences at v[3]; (3) decision situations are deterministic—no chance events occur—and are supposed to be known to the agent. That is, the agent knows in advance everything about the decision situation, the future acts available, his future preferences, and so forth; he knows everything except, perhaps, what he will actually do.

Given a choice function C one can define a binary relation of preference $x > y$ (x is preferred to y):

$$x > y \text{ if and only if } x \in C(\{x, y\}) \text{ and } y \notin C(\{x, y\})$$

III

States are valued for various reasons, and so far no assumption has been made concerning the dimensions of value associated with a state. The notion of *payment* will play an important role, however. A state u determines the resources one has for payment, $M(u)$, which I will take to be a (finite) set of some commodity that can fill the purposes of payment (for simplicity assume that we are speaking of money). I will make the minimal assumption that for any end state u and for any $\varepsilon \subseteq M(u)$, there is a state u' that, intuitively, is just like u except that $M(u') = M(u) - \varepsilon$.[4] That is for any end state u and any payment ε available at that state, there is a state u' just like u except that one has paid (an additional) ε to get there. To simplify notation I will denote such a state u' by $u - \varepsilon$ and assume that it singles out a unique state.

[3] The distinction I wish to emphasize is similar to the distinction between now-for-then/then-for-then preferences championed by R. M. Hare, *Moral Thinking* (New York: Oxford, 1988).

[4] The phrase "just like" will not be made more precise, the formal substance of this notion is provided by PNP and PPA.

(PNP) One never considers it optimal to pay for something one can have for free.

This *Principle of No Payment* (PNP) will be assumed, formally:

(PNP) For any preferences C, any set of outcomes X, any state $v \in X$ and any nonempty payment $\varepsilon \subseteq M(u)$: $C(X \cup \{v - \varepsilon\}) = C(X)$.

Let me emphasize that PNP does not imply the following principle:

$$\text{If } u > v \text{ and } u \in X, \text{ then } v \notin C(X).$$

If such a condition is a true test of coherence (it is not, not without further qualifications, or so I will argue), this is the kind of thing one would wish to prove with a pragmatic argument. The agent is at u willing to *Pay for the Preferred Alternatives* (PPA) in (the set of states) X whenever:

(PPA) For any $Y \subseteq X$, there is a nonzero payment ε such that $x \in C(Y)$ if and only if $x - \varepsilon \in C(Y^*)$, where we get Y^* by replacing each optimal element x of Y (that is, each $x \in C(Y)$) by $x - \varepsilon$.

For instance if $u > v$, then PPA is satisfied in the set $\{u, v\}$ if there is some ε such that $u - \varepsilon > v$. Note that Y^* need not be a subset of X. PPA does not imply that payment is infinitely divisible. Note also that PPA is something that *may* hold for a given set of alternatives X; I do not assume that it holds in general. I can think of three reasons why the principle of paying for a preferred alternative would not hold: (a) the agent has no money, (b) the difference in value between the alternatives in X is extremely small (at least in relationship to the smallest unit of money one can pay), or (c) the decision situation is such that introducing the notion of paying (or getting paid) for an alternative is not admissible. I will return to (b) and (c) below because I think that there are interesting cases where they truly block pragmatic arguments that have appeared in the literature.

IV

The principle to be made precise is: your preferences are incoherent if you judge it bad to follow them. As is often the case with formal explications a number of simplifying assumptions must be made and the proposed explication is but one of several more or less reasonable alternatives.

As a first step I introduce an auxiliary concept:

The agent at a state u, *considers it bad to follow plan* $p = \langle u_1, \ldots, u_n \rangle$ *through*, if there is some state $w = u_m$ such that $p(\#) \notin C_m(P_m(\#))$.

That is, the agent, at u, considers it bad to follow plan p through if there is some future node w such that the agent at u does not consider

the final outcome of p to be optimal among the alternatives available at w: plan p is not, from the perspective of u, worth following through. Note that this involves the agent's preferences at u about some future state w, no mention is made of the preferences at w.

> The agent's preferences C on states are *weakly coherent* if and only if for every non-final state u, there is some plan p such that $p(\#) \in C_u(P_u(\#))$ and the agent does not consider it bad to follow p through.
>
> The agent's preferences C on states are *strongly coherent* if and only if for every non-final state u and for every plan p such that $p(\#) \in C_u(P_u(\#))$, the agent does not consider it bad to follow p through.

An agent who violates the principle of weak coherence is *strongly incoherent* (every plan one now considers optimal is such that one (now) does not regard it worth following through), while an agent who violates the principle of strong coherence is *weakly incoherent*.

Theorem 1

(1) Any agent who satisfies (α_w) is weakly coherent:

$$(\alpha_w)\ (\forall X \subseteq S)(\exists x \in X)(\forall\ Y \subseteq X)\ (x \in Y \rightarrow x \in C(Y))$$

(2) An agent who is weakly coherent will satisfy (α_w) on any set of alternatives W that satisfy PPA (here the quantifiers in the statement of (α_w) are restricted to W).

(3) Any agent who satisfies α is strongly coherent:

$$(\alpha)\ (\forall X, Y \subseteq S)\ (C(X) \cap Y \subseteq C(X \cap Y))$$

(4) Any agent who is strongly coherent will satisfy α on any set of alternatives X that satisfy PPA.

In the next section I will consider some concrete examples. For now consider an agent with cyclical preferences over the alternatives a, b, and c: $a > b$, $b > c$, and $c > a$. Cyclical preferences are defined across binary choices but these need not determine what the agent considers optimal if he faces all three alternatives at once. As, by assumption, $C(a, b, c)$ is nonempty, the preferences are *benignly cyclical*.[5] For now, assume that $C(a, b, c) = \{a, b, c\}$—that is, when facing all three alternatives each is equally attractive.

Say that the agent's preferences over a, b, and c satisfy PPA. Consider the following decision situation (illustrated in Figure 1): the agent can choose either a, b or c; if a is chosen the agent will be offered to

[5] This distinction is due to Wlodek Rabinowicz, "Money Pump with Foresight" in Michael Almeida, ed., *Imperceptible Harms and Benefits* (Boston: Kluwer, 2000), pp. 123–52.

switch a for $c - \$1$; if b is chosen the agent will be offered to switch b for $a - \$1$; finally, if c is chosen, then the agent will be offered to switch c for $b - \$1$. The agent knows what is going to happen, thus at the outset he knows that he can achieve any one of the outcomes $a, b, c, a - \$1, b - \$1, c - \$1$. By PNP, $C(a, b, c, a - \$1, b - \$1, c - \$1) = \{a, b, c\}$.

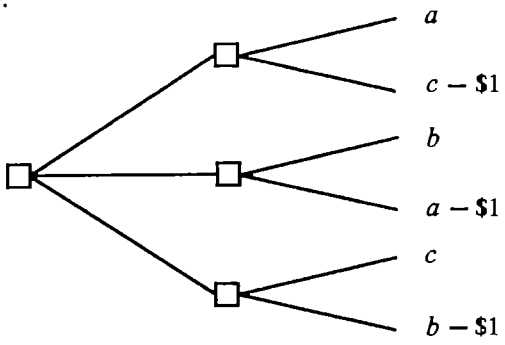


Figure 1

Consider the plan to first take a and then turn down the offer to switch a for $c - \$1$. This plan is *optimal* because its final outcome is optimal among all the available outcomes. But after embarking on this plan, the agent faces the alternatives of either keeping a or switching to $c - \$1$. By PPA, the latter alternative is preferred. Hence at this point the plan to take a is beaten by the suboptimal plan to take $c - \$1$: the agent does not consider every action recommended by the plan to be optimal among its alternatives. Indeed every optimal plan is beaten by a suboptimal plan. Therefore, the agent is (strongly) incoherent.

V

Is the incoherent agent a straw man? Is, for instance, the idea of cyclical preferences so absurd that its rejection need not be argued for in terms of pragmatic arguments? Perhaps not. In this section, I want to argue that cyclical preferences can be perfectly rational as long as the preferences are not incoherent. As cyclical preferences are incoherent when PPA is satisfied, I will focus on situations where it is not obvious that PPA should be satisfied.

There are several examples in the literature that exploit the fact that imperceptible differences between outcomes can add up to make one's preferences cyclical. I think that most of these can be handled by the fact that PPA need not be satisfied when the differences between outcomes are very small. Here I shall instead consider a decision principle that violates α and preference transitivity.

Consider in somewhat simplified form an employment practice

used at some Swedish universities. When choosing among candidates for a professorship, "equal weight" should be placed on research and teaching merits. If two candidates are otherwise considered equal the candidate (if any) who belongs to an underrepresented sex should be chosen.[6]

Let us say that there are three candidates for such a job: David, Paul, and Karen. David has both better research merits and teaching merits than Karen, and better research merits than Paul. Paul has better teaching merits than David, but worse research merits than Karen. Karen is the only candidate who belongs to an underrepresented sex. As David dominates over Karen in both disciplines, Karen cannot be considered an optimal candidate, so $C(d, k) = \{d\}$ and $k \notin C(d, p, k)$. Neither David nor Paul, however, is determinately better than the other, so $C(d, k) = \{d, k\}$. Thus in a three-way choice, David and Paul would appear to be the optimal candidates, so $C(d, p, k) = \{d, k\}$. In a two way choice between Karen and Paul neither is determinately better than the other as regards the main criteria, research and teaching, but Karen belongs to an underrepresented sex, so $C(p, k) = \{k\}$. But as $C(j, d, k) = \{j, k\}$ we have a clear violation of α (and of the transitivity of strict preference).[7] So for instance if David withdraws his application then Karen and not Paul will get the job, even though he, but not she, was previously considered a top candidate.[8]

The university's decision method thus generates a choice function that violates α. Should we protest that this is weakly incoherent and insist that the university employment policy be revised? University representatives could in their defense claim that a violation of α yields incoherence only if PPA is satisfied, but introducing alternatives where payment plays a role alters the decision situation in a way that does not bear on the problem at hand.

One can interpret this in two different ways. A strong version is that purely hypothetical decision situations are never normatively relevant: every decision situation is unique and one is free to choose any decision principle that suits the situation without worrying about what would happen if the same principle was applied in another situation. But what is meant by a decision principle that "suits" a

[6] This is a *lexicographic* decision method. See, for instance, Isaac Levi, *Hard Choices* (New York: Cambridge, 1986).

[7] If we added further layers of lexicographic principles, based for instance on underrepresented minorities, David might turn out to be preferred to Paul, in which case the preferences would become cyclical.

[8] Structurally this is just a variation of an example considered in Levi

decision situation? In principle one can for each decision situation invoke a decision principle applicable only to that decision situation. But this decision theoretical analogy to extreme ethical particularism misses the same points as its progenitor, perhaps more so.

We are frequently at a loss about what to do—or we mistrust our intuitions for the case in hand—and search for principles that at the very least are not known to give unsound advice in other situations that are relevantly similar. 'Relevantly similar' is the key phrase here because we cannot at this level of generality say a lot more about it. To this extent maybe we are all forced, at some level at least, to become decision theoretical particularists. But even if we cannot in general say exactly what we mean by a "relevantly similar" decision situation, it is a fact that we perceive particular classes of decision situations to be relevantly similar and we can give reasons for so perceiving them. The force of a pragmatic argument will depend upon how deeply entrenched and widely spread this perception of relevant similarity is.

A weaker interpretation of the university representatives defense might go roughly as follows: the decision situation in question is subject to procedural constraints and while there may be hypothetical decision situations that are similar in all respects but lack the procedural constraints, these hypothetical situations are normatively irrelevant. In short, introducing the notion of payment (of any sort) might simply be disallowed (in this case for legal reasons, bribes are not allowed): paying for the preferred alternatives is just not possible within the procedural constraints of the decision situation.

Recruiting personnel is a particular kind of decision situation with particular procedural constraints (these will of course differ between different institutions and may be implicit). We may agree on the dimensions of value involved (research merits and pedagogical merits and that it is good for the university to have a roughly equal representation of the sexes) although we may have different ideas about how to amalgamate them. Whatever decision principle we invoke, however, we should be prepared to defend it against allegations of arbitrariness, incoherence, and so on. But the university's decision method, I would argue, stands up to these tests in this context: (a) the activity of recruiting professors does not seem to single out an arbitrarily chosen class of decision situations to which to restrict a decision method—after all, there are laws that specifically regulate this activity (by contrast, imagine that the procedural constraints dictated that we employ one method on Mondays and another on Thursdays)—and (b) in this class of decision situations violation of α does not yield incoher-

ence. By accepting (a) we reject PPA for this class of decision problems, and without PPA we cannot get the pragmatic argument going.

Frederic Schick[9] has claimed that cyclical preferences "are incoherent because they violate the principle of preference transitivity, not because they can be exploited" (*ibid.*, pp. 112–13). To the contrary I would claim that cyclical preferences can be perfectly coherent or rational, although only in limited contexts. Our perceptual apparatus can, arguably, give rise to cyclical preferences, but these preferences are incoherent only if an additional condition is fulfilled: that one judges that the differences are worth paying for. We cannot give a carte blanche to decision methods that give rise to cyclical preferences (or violate α), but arguing against such decision methods solely on the basis that cyclical preferences (or violations of α) are self-evidently absurd goes beyond what can be argued for in pragmatic terms.

Of course, one can argue, just as I must argue, that we must start somewhere, we must make some assumptions. Why not start with acyclicity or α? Well, turning to the university employment practice, there is the very real question why a violation of α would be a reason not to use the practice. I have argued that the violation of α in this case has no pragmatic repercussions whatsoever, thus there seems to be no substantial reason to abandon the employment practice on purely structural or formal grounds. Violation of "obvious" constraints on decision methods, such as acyclicity and α, can be reasonable. But as soon as we start looking at situations where the differences between alternatives are not minuscule and where procedural constraints do not enter into play, violation of acyclicity or α should be considered bad. Why? Not because such a violation is self-evidently absurd, but because in these situations the preferences would be incoherent.

VI

So far, only preferences on outcomes have been considered, but it has already been suggested that these are related to preferences on plans and actions. Here I shall simply assume that the agent is not myopic, and that preferences for actions are derived from preferences on plans according to the principle that the best action is the initial action prescribed by the best plan:

(PA) For any states u, v, w: $w \in C_u(A(v))$ if and only if for some $p \in C_u(P_u)$, $p(1) = w$.

While reasonable, PA is not true by conceptual reasons alone. One

[9] "Dutch Bookies and Money Pumps," this JOURNAL, LXXXIII, 2 (February 1986): 112–19

can think of situations where one considers a plan optimal, but cannot bring oneself to perform the first action in the plan. An example might be where the plan is to quit smoking and the first action dictated by the plan is not to smoke the cigarette one has in one's hand. Preferring to smoke the cigarette might be irrational, but it surely is not a conceptual impossibility.

A plan that one knows can be followed through if adopted is *feasible*. In the present setting the only thing that can influence whether the agent will follow through a plan is the future preferences of the agent. Thus, because of PA a plan p is feasible if and only if at every future continuation of p the plan is at that point considered optimal.[10] We can define this recursively. The feasible plans at u, in symbols F_u, is the smallest set satisfying:

(1) If $p = \langle v \rangle$ is a plan for u, then $p \in F_u$.
(2) If $p = \langle w, v_1, \ldots, v_n \rangle$ is a plan for u and $v_n \in C_w(F_w(\#))$, then $p \in F_u$.

When X is a set of plans let X_F denote the feasible plans in X. An agent who believes that he at each future state will follow his preferences at that state will believe that an infeasible plan will not be followed through. Thus while he may consider it optimal to follow an infeasible plan through, he believes that this will not happen. An infeasible plan is, in a distinct sense, not available to the agent. The action sequence described by an infeasible plan is still, of course, a possible action sequence (in the sense that each action would be available to the agent), it is just that at some point his future self (if he acts according to his future preferences) will *choose* to do some other action than the one prescribed by the plan.

A simple example will illustrate. Take Ulysses's decision problem before facing the Sirens. He can choose to take no precautions, in which case when the time comes he will heed the call of the Sirens and go to his doom. Or he can tie himself to the mast, hear the Sirens and survive. Obviously the latter is the prudent course of action. But there is a plan that has a still better outcome if it is followed through: take no precautions yet do not, when the time comes, jump ship. The problem with this plan is, of course, that it is not feasible. Ulysses knows that by the time he hears the Sirens he will want to jump ship, he knows that his preferences will change, hence the plan is suboptimal, even though its outcome, if followed through, would be optimal.

[10] A detailed exposition of feasibility can be found in P. J. Hammond, "Changing Tastes and Coherent Dynamic Choice," *Review of Economic Studies*, XLIII (1976). 159–73.

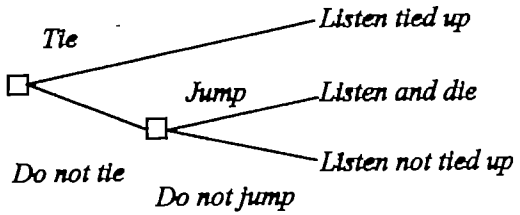


Figure 2

For this reason one should resist the temptation to derive the value of plans from the value of their outcomes if followed through (their end states) according to the following principle:

(OP°) For any state u and set of plans X:

$$p \in C_u(X) \text{ if and only if } p(\#) \in C_u(X(\#))$$

Even though the end state of a plan is assumed to be the "prime conveyor of value" for the plan, it is not the only feature that determines the value of the plan. Rather, the principle we are after is the following:

(OP) For any state u and for any set of plans X:

$$p \in C_u(X) \text{ if and only if } p(\#) \in C_u(X_F(\#))$$

Theorem 2

(1) If the agent's preferences on states are strongly coherent and either OP° or OP is satisfied, then for any states u and v:

$$C_u(P_v) \cap F_v \subseteq C_u(F_v)$$

(2) If the agent's preferences on states are strongly coherent and satisfies OP, then for any states u and v:

$$C_u(P_v) = C_u(F_v)$$

There is a *pragmatic* justification for rejecting OP° in terms of coherence.

Say that you adhere to OP1° and PA. This would mean that if p is an infeasible plan that is considered optimal (that is, $p \in C_u(P_v)$ but $p \notin F_v$), the action $p(1)$ will be considered optimal (because of PA). But if this first action does not coincide with the first action proposed by some other plan $q \in C_u(F_v)$, you can, by performing $p(1)$, no longer reach any of the outcomes in $C_u(F_v(\#))$. Once $p(1)$ has been performed there is no turning back. Of course even if p is not feasible, some other action sequence beginning with the action $p(1)$ must be feasible. Hence the action sequence that will actually be followed through if you perform the first action prescribed by p will

correspond to some feasible plan q such that $q \notin C_a(F_a)$. By Theorem 2, $C_a(P_a) \cap F_a \subseteq C_a(F_a)$. As $q \in F_a$ and $q \notin C_a(F_a)$, it follows that $q \notin C_a(P_a)$. Hence you know that if you were to follow your present preferences for plans, you would end up following an action sequence (indeed a plan) that you presently consider suboptimal: your preferences are incoherent.

Of course one may argue that this whole derivation is unnecessary, as, by conceptual reasons alone, no rational agent can intend to follow through a plan that he knows he will not follow through. But this is going too fast: there need be no intention to follow the plan through. The adherent of OP° could for instance reason as follows: (1) the action sequence $p = < a, b, c >$ would if followed through have the outcome r; (2) r is the best outcome that could be reached by any sequence of actions, hence p is the best action sequence; (3) as p is the best action sequence and it prescribes a as the first action, a is the best action at this point; hence (4) I will do a. This way of reasoning would be a way of deriving a preference for actions that results in the intentional action a. The derivation of the value of a, however, makes no mention of any intention to follow through the plan p.

The idea of adopting a plan, of intending to follow it through, is obviously an important concept in the theory of action, but one can act intentionally without having an intention to follow through a particular multi-action plan. Of course if one derives the value of an action by appealing to an infeasible plan, one can justly be criticized. But not for violating some conceptual constraint on the notion of preference for plans or of intentional action, but because of incoherence: one has a reason to judge it bad to follow this preference on action, hence this way of deriving that preference is flawed.

The original definition of coherence in section IV applied to preferences on *outcomes* but utilized the notion of a plan. The idea, roughly, was that preferences on outcomes are incoherent if there is a plan p such that the outcome of p, if p were followed through (that is, $p(\#)$), was considered optimal but one at the same time did not consider it optimal to follow p through. The key phrase here is the outcome of a plan *if it were followed through*. No mention is made of the feasibility of a plan as the notion of feasibility always involves *future* preferences and the coherence of one's present preferences does not depend on future preferences.

Consider again the Ulysses example. His preferences over the outcomes are unproblematic (coherent) as they are transitive: *listen-and-die* is worse than *listen-tied-up* which is worse than *listen-not-tied-up*. His preference for plans is also unproblematic (coherent) as it complies with OP: the optimal plan to follow is the plan that results in *listen-*

tied-up as, of the two outcomes that can be reached with a feasible plan, *listen-tied-up* is the best. There is no contradiction in saying that although there is a plan that if followed through would give a better result (namely *listen-not-tied-up*), adopting that plan is worse (as it will not be followed through).

Consider instead the example of cyclical preferences as was illustrated in Figure 1. Recall that the optimal outcomes were *a*, *b*, and *c*, but the agent considered $a - \$1$ better than *b*, $b - \$1$ better than *c*, and $c - \$1$ better than *a*. Hence while he, for instance, considered the outcome of the plan to take *a* if followed through to be optimal and the plan to take $b - \$1$ if followed through to be suboptimal, he also considered that it would be better to take $b - \$1$ than *a* if those were the only options.

Assume now that his preferences endure, so that if he were to embark on the plan to take *a*, he would then prefer to take $b - \$1$ (in R. M. Hare's terminology, this would be the agent's then-for-then preference). This assumption was not necessary to establish incoherence as incoherence when applied to preferences on outcomes only involving now-for-then preferences, but it is a reasonable assumption. In this case the plan to take *a* is infeasible. Indeed the only feasible plans are those with $a - \$1$, $b - \$1$, and $c - \$1$, respectively, as outcomes. By OP, these plans and not the plans with outcomes *a*, *b*, and *c* should be considered optimal. But on my account the agent is already incoherent due to his preferences on outcomes, and adhering to OP at this stage will not change this fact.

So according to the notion of coherence I have been proposing here, Ulysses is perfectly coherent, but the agent with the cyclical preferences is incoherent. This is controversial. Some authors have argued to the effect that both Ulysses and the agent with cyclical preferences are coherent, and some, indeed, that they are both *in*coherent. Buried herein lies what I take to be one of the most fundamental controversies as regards the validity of pragmatic arguments.

VII

The *Principle of Reflection* states the following relationship between the probabilities one assigns now at time 1 (P) and the probabilities that one will assign at some future time 2 (Q):

For any proposition A and number x, $(0 \leq x \leq 1)$: $P(A \mid Q(A) = x) = x$

As Bas C. van Fraassen has shown, if one violates Reflection one can be exploited.[11] Consider an example. Let A be some future event and let E be the future event that $Q(A) = .5$. Assume that $P(E) = .2$.

[11] "Belief and the Will," this JOURNAL, LXXXI, 5 (May 1984): 235–56.

Let $P(A \mid E) = .75$, that is, the agent violates Reflection. At time 1 the agent is offered the following two bets that he must either jointly accept or reject:

(bet 1) Pay \$0.50 to win \$3.60, if E occurs.
(bet 2) Pay \$15.00 to win \$20.20, if A occurs; this is conditional on the event E. If E does not occur then the \$15.00 will be refunded.

Now, if E occurs at time 1, the agent will be offered the following bet:

(bet 3) Pay \$10.00 to win \$20.20, if A does not occur.

In Figure 3, we see the structure of the decision problem and the payoffs for the various contingencies.

There are four plans available. Working out the expected monetary value from the perspective of the agent's preferences at time 1, we see that the plan (a) <accept bets 1+2 and if E occurs reject bet 3 > is the plan which if followed through has the optimal expected value, followed by (b) <reject bets 1+2 and if E occurs reject bet 3> which is better than (c) <accept bets 1+2 and if E occurs accept bet 3>. The plan (d) <reject bets 1+2 and if E occurs accept bet 3 > has the worst expected monetary outcome. Of course if E occurs, then from the perspective of the preferences of time 2, <accept bet 3> is better than <reject bet 3> regardless of whether one previously has accepted bets 1+2 or not. Hence, of the four plans originally available only (c) and (d) are feasible. As the expected value of (c) is greater than (d), the agent should according to OP initially prefer the plan (c) to (d) and hence by PA accept the bets 1 and 2.

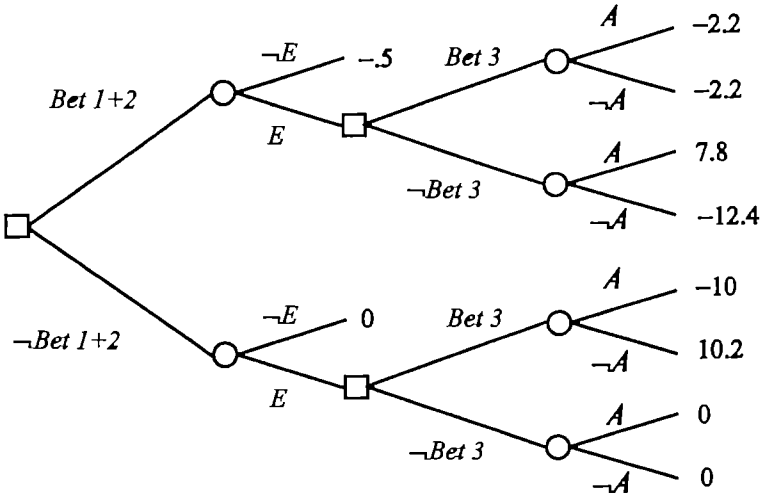


Figure 3

The problem with accepting bets 1 and 2 is that if E does not occur one loses money, while if E occurs one will accept bet 3 and one will again lose money regardless of whether A is true or not. Thus by accepting the two first bets one exposes oneself to a sure monetary loss. In this sense one has been *Dutch booked* and in this sense one has been exploited. Of course, while accepting bets 1 and 2 gives a *sure* loss, it also gives the least *expected* loss (from the perspective of time 1).[12]

On my account, the agent is coherent; indeed, the structure of the problem is essentially that of Ulysses and the Sirens: the preferences for outcomes are transitive (hence coherent) and the preferences for plans are derived by OP (hence coherent). Thus a violation of Reflection need not be due to incoherence and so an agent who is exploitable need not be irrational. This last conclusion has been reached by several authors on different grounds.

A number of authors have argued that there are clear counterexamples to the Principle of Reflection.[13] As soon as one suspects that one might in the future become irrational, be it because of the fact that insanity runs in the family, or because one has just ingested a mind-bending drug, one will violate Reflection. A suspicion of future irrationality should by itself not be considered irrational, yet even if the suspicion turns false so that one never becomes irrational, the fact that one violates Reflection makes one exploitable. Hence no irrationality at all may befall the agent, yet he can violate Reflection and become exploitable.

Isaac Levi[14] has objected to Reflection on epistemological grounds. A consequence of Reflection is that one would be irrational if one considered it possible that one in the future revise one's belief in any proposition that one currently accepts. For if $P(A) = 1$, then $P(A \mid Q(A) = x) = 1$, for whatever value of x one chooses. Hence if one does not now consider the falsity of A a serious possibility, one is by Reflection committed to not considering the falsity of A a serious possibility in the future as well. Thus Reflection imposes a rigidity that Levi finds unacceptable.

Even if many do not find Reflection to be a true principle of rationality, the fact that a violation of Reflection can be exploited by a Dutch book argument has led people to draw different conclusions.

[12] This is essentially the line of argument presented by Brian Skyrms, "A Mistake in Dynamic Coherence Arguments?" *Philosophy of Science*, LX, 2 (June 1993): 320–28.

[13] See, for instance, David Christensen, "Clever Bookies and Coherent Beliefs," *The Philosophical Review*, C, 2 (April 1991): 229–47.

[14] "The Demons of Decision," *The Monist*, LXX, 2 (1987): 193–211.

Some have argued that this shows that there is something wrong with pragmatic arguments, the argument form is just not valid: Dutch books and money pumps establish nothing as regards irrationality.

But Levi, notably, has argued that the problematic feature of the Dutch book argument for Reflection is not shared by the synchronic Dutch book arguments proposed by Frank P. Ramsey and Bruno de Finetti and so it is not the argument form as such that is put into question, but the way it is used to defend Reflection (*ibid.*).[15] Levi argues that in the case of Reflection one is, in fact, not being exploited: of the feasible options, one is choosing the best and exploitation only occurs when one's choice is dominated by another feasible option. There is a minor terminological difference here as I would say that one is being exploited as soon as one suffers a sure loss, but Levi and I concur on the judgment that one need not be irrational when this occurs.

Levi's diagnosis why exploitability (in my sense of the term) does not entail irrationality has some similarities to the view I have proposed here. The plan that, if followed through, would have the optimal outcome is not feasible; hence, one knows that it will not be followed through, and so, according to Levi, it is not an option. As the perspective adopted by Levi is that of adopting a plan I would agree: for a rational agent, the option of adopting an infeasible plan is not an option. But I would go further and say that even if the agent is not adopting a whole plan, he should not, if rational, use the outcome of an infeasible action sequence to derive the value of the first action in the action sequence.

These are minor subtleties. The main difference between Levi's view and mine becomes clear if we once again return to the case of cyclical preferences as illustrated in Figure 1. We agree that if the agent's preferences endure, the plans to take a, b, or c, respectively, are infeasible. But according to Levi, this means that the agent is not incoherent even though the agent will end up with one of the outcomes $a - \$1$, $b - \$1$, or $c - \$1$, because, Levi argues, these are the only outcomes he can reach (if he follows his preferences). Thus while Levi and I agree that violation of Reflection is not a sign of irrationality, we disagree about whether cyclical preferences in the presence of PPA should be considered irrational. On my view, the quest for feasibility is derived from the quest for coherence, but

[15] A similar point is made in Rabinowicz, "On Seidenfeld's Criticism of Sophisticated Violations of the Independence Axiom," *Theory and Decision*, XLIII (1997): 279–92.

adopting a feasible plan when one's preferences for outcomes are incoherent does not make one coherent. On Levi's view one's preferences for outcomes can only be judged incoherent in relation to the outcomes that can be reached by a feasible plan.

Let me give some prima facie reasons in favor of my position. First, the agent with cyclical preferences will end up paying for an option that he could have had for free, 'could' in the sense that it will be an option that he would choose not to take. And specifically, he does this even though his preferences endure. Both in the case of Ulysses and in the case of Reflection the agent's preferences change over time and this is the reason that the plan with the optimal outcome is not feasible. The agent's preferences at a particular time cannot be held to be incoherent for not reaching an optimal outcome when the preferences change. But when the preferences endure and are still not able to deliver an optimal outcome, this arguably is a prima facie case for criticizing the preferences.

Second, when we are just considering preferences on outcomes, feasibility is a contingent property. Say that we embellish the example so that the agent can choose to go ahead with his present preferences but also has the choice of reversing his preferences so that b is better than $a - \$1$, c is better than $b - \$1$, and a is better than $c - \$1$. So, for instance, he can choose to face the option between b and $a - \$1$ either with his present preferences or with reversed preferences. If he satisfies OP, he will prefer to enter into the decision situation with reversed preferences (as that will give him b rather than $a - \$1$). So to get an optimal outcome (according to his present preferences) he must choose to reverse his preferences; he does not find his present preferences a good guide for action. Again, I would say, this is a strong prima facie case for saying that his preferences are not coherent.

Of course, one might concede that there is something "odd" about having incoherent preferences (in the sense I am proposing), yet maintain that to the extent that incoherence appeals to outcomes that can only be reached by an infeasible plan, incoherence is not a *pragmatic* notion (and perhaps, to this extent, has no normative force). After all, one might say, if we are strictly pragmatic we should only appeal to outcomes that the agent can actually reach when evaluating the performance of the agent. But this to some extent would be begging the question. The issue is in part what exactly should count as pragmatically problematic behavior and in part how the blame for this should be apportioned. I would say that the very fact that the agent with cyclical preferences ends up paying for something he could have had for free is a real pragmatic problem for the agent. Indeed, the fact that a violation of Reflection can lead to exploitation is *also*

a pragmatic problem for the agent. The difference between the two cases, in my view, is that the source of the latter pragmatic problem is not the initial preferences of the agent and the solution to it is not for the agent to revise these preferences, but rather, if anything, to buy an insurance policy (accept bets 1 and 2) so as to reduce the damage done if his preferences change. Levi, by contrast, would find neither case pragmatically problematic for the agent.

So the bone of contention lies not in what the agent can achieve, but in what should count as a pragmatic evaluation of the agent's achievements. As this is done in the light of the agent's own preferences one should, I argue, look at the problems those very same preferences generate. If they themselves make the optimal outcomes infeasible, this is a pragmatic problem. The implicit analogy between this issue and 'ought implies can' is misleading because what the agent with enduring preferences *prefers* to do determines both what he *can* do and what he *ought* to do if we judge him only by his own preferences and it is what he prefers to do that is at issue. The rational agent should only embark on feasible plans, but this is a constraint derived from the assumption that the agent is coherent and not from the principle that outcomes that cannot be reached ('cannot' in the sense that they are not the outcomes of any feasible plan) are inadmissible as objects of comparison when evaluating the agent's performance.

To conclude: an agent who is coherent is not exploitable, *if his preferences endure*. An agent who is incoherent is exploitable, *if his preferences endure*. This links coherence to performance. But when preferences do not endure, exploitability is neither a necessary nor sufficient condition for incoherence. This weakens the link between coherence and performance, but only because it is a contingent matter whether or not one's preferences endure, just as it is a contingent matter whether or not one will ever end up in a decision situation where one's preferences, incoherent or not, are put to the test.

JOHN CANTWELL

Royal Institute of Technology/Sweden

PROOFS OF THE THEOREMS

Proof of Theorem 1. (The proofs of 1 and 3 are similar in character as are 2 and 4 so here I will prove only 3 and 4.) (3) Assume that the agent is standard and satisfies α. Note that if v is reachable from u, then $P_v(\#) \subseteq P_u(\#)$. By α, if $x \in C(P_u(\#))$ and $x \in P_v(\#)$, then $x \in C(P_v(\#))$. Hence no optimal plan at u is considered bad to follow through.

(4) Assume that the agent is strongly coherent, and assume that X and Y are finite sets of end states, that $u \in C(X) \cap Y$ and for *reductio* that $u \notin C(X \cap Y)$. Let Z^* be the set we get by replacing each optimal element x of $X \cap Y$ by $x-\varepsilon$ for some suitable (nonzero) ε. Let v be a state where $A(v) = X - \{u\}$ and let w be a state such that $A(w) = Z^*$. Let m be a state where $A(m) = \{v, w\}$.

Thus at m the agent can choose between the state v (which allows a choice between the elements of $X - \{u\}$) and the state w (which allows a choice between the elements of Z^*). As $u \in Z^*$ (remember u is not an optimal element of $X \cap Y$) the final outcomes that can be reached from m are $X \cup Z^*$. Also as $u \in Z^*$, the plan $p = \langle w, u \rangle$ is a plan for m, furthermore $p(\#) \in C(P_m(\#))$ (as $p(\#) = u$ and $P_m(\#) = X \cup Z^*$). By PPA, as $p(\#) \notin C(X \cap Y)$, $p(\#) \notin C(Z^*)$, thus $p(\#) \notin C(P_m(\#))$. Hence p is an optimal plan at m that is not worth following through; hence, contrary to assumption, the agent is not strongly coherent.

Proof of Theorem 2. (1) Assume that the agent's preferences on states are strongly coherent. Thus in particular, for any u and v, $C_u(P_v(\#)) \cap F_v(\#) \subseteq C_u(F_v(\#))$. Assume now that $p \in C_u(P_v) \cap F_v$. Assume that the agent satisfies OP°, then $p(\#) \in C_u(P_v(\#))$. As $p \in F_v$, $p(\#) \in F_v(\#)$, so $p(\#) \in C_u(F_v(\#))$. Again by OP°, $p \in C_u(F_v)$. Assume instead that the agent satisfies OP. Thus, as $p \in C_u(P_v)$, $p(\#) \in C_u(F_v(\#))$. Then by OP, $p \in C_u(F_v)$.

(2) Assume that the agent's preferences on states are strongly coherent and that he satisfies OP. In the proof of (1) we saw that $C_u(P_v) \subseteq C_u(F_v)$. To show inclusion in the other direction, assume that $p \in C_u(F_v)$. By OP, $p(\#) \in C_u(F_v(\#))$, but then, again by OP, $p \in C_u(P_v)$.

FROM CONSEQUENTIALISM TO UTILITARIANISM*

In this article, we show that *total act utilitarianism* can be derived from a set of axioms that are (or ought to be) acceptable for anyone subscribing to the basic ideals of consequentialism.

I. INTRODUCTION

Total act utilitarianism prescribes that an act is morally right just in case the total sum of utility for all individuals is at least as high as that of all alternatives. The term 'consequentialism' denotes a more general set of ethical theories, according to which the moral status of an act is entirely determined by its consequences. Derek Parfit's *priority view*[1] is an example of a nonutilitarian version of consequentialism, which prescribes that benefits to the worse off should count for more than benefits to the better off. Another example is the *maximin principle*, which entails that an act is morally right just in case the consequences for the worst off individual are as good as possible.[2] Further examples include different versions of egalitarianism, prescribing that the relative differences in well-being among individuals are ethically important (in most cases together with the total amount of well-being).

A plausible point of departure for a defense of total act utilitarianism is the question: Why is total act utilitarianism better than other versions of consequentialism, such as the priority view, the maximin principle, and different forms of egalitarianism? This is equivalent to asking how we know that utilitarianism is correct *given* that some version of consequentialism is known to be correct. Needless to say, a satisfactory answer to this question does not resolve all ethical problems. A substantial gap between consequentialism on one hand, and rights, duty, and virtue ethics, on the other hand, will remain.

Unlike John C. Harsanyi's famous utilitarian theorem and John Broome's improved version of this theorem,[3] the axiomatization pre-

*I wish to thank Sven Ove Hansson, John Cantwell, Gunnar Bjornsson, Hans Mathlein, Torbjorn Tannsjo, and an anonymous referee for valuable comments on earlier drafts.

[1] "Equality or Priority?" The Lindley Lecture, 1991 (Lawrence: University of Kansas, 1995); see also a shortened version, "Equality and Priority," *Ratio*, X, 3 (1997): 202–21.

[2] The most well-known application of the maximin principle to an ethical context is John Rawls's *A Theory of Justice*, (Cambridge: Harvard, 1971).

[3] See, for example, Harsanyi, "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," *Journal of Political Economy*, LXIII (1955): 6–23, and "Bayesian Decision Theory, Rule Utilitarianism, and Arrow's Impossibility Theory" *Theory and Decision*, XI (1979): 289–317—see also Broome's *Ethics Out of Economics* (New York: Cambridge, 1999), chapter 6.

0022-362X/03/0008/403–415 © 2003 The Journal of Philosophy, Inc.

sented here does not rely on Bayesian decision theory. Hence, the new axiomatization will appeal also to people who do not find the axioms of Bayesian decision theory acceptable.

A difficulty in Harsanyi's and Broome's Bayesian approaches, which has been recently discussed by Wlodek Rabinowicz, is that their axioms only imply that, "The overall goodness is an additively separable function of the individual goodness values, but they do not guarantee that each individual is treated equally in this addition."[4] In order to get all the way to utilitarianism, Harsanyi and Broome also have to assume that we somehow can calibrate the utility functions of different individuals and thereby make sure that everyone's benefits count equally much. This would, obviously, presuppose cardinal comparisons of interpersonal utility. The axiomatization presented here does not solve the problem of interpersonal utility comparisons, but *cardinal* comparisons of interpersonal utility play a more salient role in the new axiomatization, compared to Harsanyi's.

In section II, a set of ethical axioms are proposed, and in section III, two utilitarian theorems are stated. The proofs of the theorems are given at the end.

II. ETHICAL AXIOMS

Somewhat roughly put, we are going to derive utilitarianism from the following five consequentialistic intuitions:

(1) (The value of) consequences form an extensive structure.
(2) In every ethical problem, at least one act is morally right.
(3) If an act is split into two other acts, which both have exactly the same consequences as the original act, then the two new acts are morally right just in case the original act is morally right.
(4) If an act yields strictly better consequences for all individuals than another, then the latter act is not morally right.
(5) If the consequences of an act are better for individual i than for i', then there is some nonzero amount by which the consequences for i can be diminished, such that there is some (possibly large) increase in the consequences for i' that compensates for this.

We shall now render these consequentialistic intuitions more precise. Let $A = \{a, a', a'', \ldots\}$ be a nonempty and finite set of alternative acts. We conceive of acts as particular acts rather than as act-types, and naturally all alternative acts should be performable by the same agent.[5]

[4] "Prioritarianism for Prospects," *Utilitas*, XIV (2002): 2–21, see p. 6.

[5] It should be emphasized that acts are choices about what total amount of good and bad consequences each individual will experience from *now* until he or she *dies*. That is, by choosing an act we choose each individual's *final* amount of good and bad consequences.

Furthermore, let $I = \{i, i', i'', ...\}$ be a nonempty and finite set of individuals. In an ethical context such as this, it is reasonable to assume that the set of individuals includes all individuals in the world—not just some specific subset, such as white South Africans. For each pair of acts and individuals in $A \times I$ there are consequences c_{ai}, $c_{ai'}$,.... The set of all possible consequences individuals may face is denoted C. It is not concrete states of affairs that matter from an ethical point of view for consequentialists (at least not according to the versions considered here). Instead, a 'consequence' should rather be conceived of as the entity that the consequentialist claims to be valuable, for example, preference satisfaction or amounts of hedonistic pleasure.

Causal and other relational aspects of consequences will not be modeled. The following example illustrates this point: suppose that the consequence *no enemy soldiers killed* is preferred to the consequence *many enemy soldiers killed*. Then, if no causal or other relational aspects are modeled, it will be reasonable to also assume that the consequence *no enemy soldiers killed & war is lost* is preferred to the consequence *many enemy soldiers killed & war is lost*, since we have added the same consequence to both initial consequences. Some (nonutilitarian) consequentialists, however, would probably consider it better that many enemy soldiers are killed in case the war is lost, which contradicts the assumption introduced above. Arguably, the best way to make sense of this noncausal and nonrelational notion of consequences is to think of consequences in a more abstract way than we usually do, for example, by focusing on their value properties, as suggested above.

In any axiomatic treatment of utilitarianism, one must at some point introduce the idea of interpersonal utility comparisons. One option is simply to assume that there is a cardinal function assigning real numbers to all consequences. This would not be entirely unreasonable, since adherents of different versions of consequentialism disagree about which rule should be used for evaluating a given distribution of consequences, not about the cardinal function itself. In the present exposition, however, we shall take another approach, and instead start from the assumption that the agent is able to make ordinal interpersonal comparisons.[6] More precisely, we assume that the agent's preferences among the elements of C form a weak ordering. This means, among other things, that a consequence c_{ai} either is better for i (is preferred to) than consequence $c_{a'i'}$ is for i', or of

[6] In the end, this assumption also implies that cardinal comparisons of interpersonal utility are possible, as will soon become evident, so this assumption is no radical improvement over Harsanyi's.

equal value, or worse. Axiom 1 (extensive structure) can then be applied for constructing a cardinal utility function.[7] An *ethical problem* can be conceived of as a triple $\delta = \langle A, I, C_{AI} \rangle$, in which A and I are defined as above and C_{AI} is the set of consequences corresponding to $A \times I$, that is, all consequences that are obtainable for the individuals I with the acts A. We then define a *solution* to an ethical problem as follows:

Definition 1. Let $\delta = \langle A, I, C_{AI} \rangle$. Then, $\Phi(\delta)$ is a *solution* to δ if and only if
(i) $\Phi(\delta) \subseteq A$, and
(ii) every member in $\Phi(\delta)$ is morally right to perform.

The solution $\Phi(\delta)$ is, of course, not supposed to be known in advance by the agent. If anything, it is the aim of normative ethics to determine the members of $\Phi(\delta)$. The triple $\langle C, \succ, \sim \rangle$ is a comparison structure for consequences, in which $\succ$ and $\sim$ are relations on C representing strict preference respectively indifference.[8] (As usual, $c \succcurlyeq c' \Leftrightarrow c \succ c' \vee c \sim c'$.) We assume that there is a binary operation $\cdot$ on C. Intuitively, $c \cdot c'$ means "both consequence c and c' obtain." Furthermore, we wish to model the idea of an "interval" of consequences that have a bounded (small) negative value, for example, "Anne loses ten cents," "Mark arrives one minute earlier to work than necessary," and so forth. Consider the following definition.

Definition 2. Let $D (\subset C)$ be a set of *bounded undesirable* consequences just in case there are some $c, c' \in C$ such that
(i) for every $d \in D$ it holds that $c \succ d \succ c'$, and
(ii) for every $c'' \in C$ and every $d \in D$, if $c'' \succ d$ then $c'' \succ c'' \cdot d$, and
(iii) for every $d, d' \in D$ in D there is some d'' such that $d \succ d'' \succ d'$.

Let us now consider the following ethical axioms, which I propose hold for every ethical problem δ. The consequences c, c', c'', c''' are arbitrary elements in C, and C_{ai} is the concatenation of all consequences faced by i in case a is performed (in the ethical problem $\langle A, I, C_{AI} \rangle$).

Axiom 1. (*Extensive Structure*) For all $c, c', c'', c''' \in C$:
(i) $\langle C, \succ, \sim \rangle$ is a weak order.

[7] As stated below, Axiom 1 will allow C to be infinitely large. At the expense of considerable technical complexity, however, Axiom 1 could be relaxed such that only a set C of more realistic size has to be considered. A good starting point for such a modification is to apply the results in David H. Krantz et al., in *Foundations of Measurement, Volume 1: Additive and Polynomial Representations* (New York: Academic, 1971), chapter 3.4 and 3.5.

[8] The concept of comparison structures is introduced and investigated in detail by Sven Ove Hansson in *The Structure of Values and Norms* (New York: Cambridge, 2001).

(ii) $[c \bullet (c' \bullet c'')] \sim [(c \bullet c') \bullet c'']$
(iii) $c \succ c' \Leftrightarrow (c \bullet c'') \succ (c' \bullet c'') \Leftrightarrow (c'' \bullet c) \succ (c'' \bullet c')$
(iv) If $c \succ c'$, then there is a positive integer n such that $(nc \bullet c'') \succ (nc' \bullet c'')$, where nc is defined inductively as: $1c = c$, $(n + 1)c = nc \bullet c$.

Axiom 2. (*Action Guidance*) $\Phi(\delta) \neq \varnothing$

Axiom 3. (*Split of Acts*) Suppose that $\delta = \langle A, I, C_{Ai} \rangle$ and $\delta' = \langle A', I, C_{A'i} \rangle$, with $A' = (A - \{a\}) \cup \{a', a''\}$. Then, if $c_{ai} \sim c_{a'i} \sim c_{a''i}$, for all $i \in I$, it holds that $\{a', a''\} \subseteq \Phi(\delta')$ if and only if $a \in \Phi(\delta)$.

Axiom 4. (*Weak Pareto*) If there is an a such that for some a' it holds that $c_{a'i} \succ c_{ai}$ for all i, then $a \notin \Phi(\delta)$.

Axiom 5. (*Trade-Off*) There is a set of bounded undesirable consequences D_{ai} such that for every $d_{ai} \in D_{ai}$ there is a (possibly much better) consequence c'_{ai} such that for all ethical problems δ, if $a \in \Phi(\delta)$ and $C_{ai} \succ C_{a'i}$, it holds that $\Phi(\delta) = \Phi(\delta')$, where δ' is the ethical problem obtained from δ by substituting C_{ai} with $C_{ai} \bullet d_{ai}$ and $C_{a'i}$ with $C_{a'i} \bullet c'_{a'i}$.

Axiom 1 enables us to assign real numbers to consequences representing utility. This is the only axiom that is not a necessary consequence of utilitarianism. In principle, it could be replaced by any set of assumptions that allow for a numerical representation of utility (that is unique at least up to a positive linear transformation).

Axiom 2 would, of course, not be accepted by people who believe in moral dilemmas. Consequentialists, however, typically show little sympathy to that idea, and therefore this is not a serious problem here.

Axiom 3 asserts that if an act is split into two other acts, which both have exactly the same consequences as the original act, then the two new acts are morally right just in case the original act is morally right. In order to illustrate what it means to split an act into two, suppose that you have decided to go for a walk in the park. This activity can be performed in two ways, namely by carrying the umbrella in either your left or in your right hand. If the consequences of these two new acts are the same as the consequences of the original act, then the moral status of all three acts should be the same. Axiom 3 follows from the basic consequentialistic idea that consequences alone determine the moral status of an act.

Axiom 4 is a version of the well-known Pareto principle, which plays a central role in Harsanyi's as well as in Broome's axiomatizations.[9] Note, however, that Axiom 4 is weaker than the traditional Pareto principle.

[9] Axiom *C* in Harsanyi, "Bayesian Decision Theory, Rule Utilitarianism, and Arrow's Impossibility Theory," p. 293. Broome discusses his use of the Pareto principle on p. 92 (*op. cit.*).

Axiom 5 is the most controversial of the five axioms. It asserts that a withdrawal of a small amount of utility from someone who is well off can be ethically compensated by giving *some* amount of utility (perhaps much more) to a worse-off individual. Of course, this axiom will not be accepted by someone subscribing to, for example, ethical egoism. One of Harsanyi's axioms, however, is also incompatible with ethical egoism, namely Axiom *D*, which states that one should give equal weight to everyone's utility functions (*ibid.*, p. 294). Furthermore, it is also worth noticing that ethical egoism cannot even be formulated in the technical framework used in this paper, since it does not allow us to tell which $i \in I$ is identical to the agent. Already this fact is, arguably, a partial reason for regarding ethical egoism as a pathological form of consequentialism.

It is trivial to show that Axioms 2 and 3 are necessary consequences of at least four versions of consequentialism, namely utilitarianism, prioritarianism, the maximin view, and egalitarianism.[10] Axiom 4 (Weak Pareto) is also a necessary consequence of these versions of consequentialism, with the exception of the most extreme version of egalitarianism; people who are concerned with nothing except value differences do not have to accept Axiom 4 (or Harsanyi's or Broome's stronger versions of the Pareto principle). Parfit's levelling down objection is a well-known argument to the effect that such an extreme form of egalitarianism is unreasonable (*op. cit.*, pp. 17–18). Axiom 5, or the weaker Axiom 5′ stated in the next section, is entailed by utilitarianism, prioritarianism, egalitarianism, and the maximin view, given that there is a set of bounded undesirable consequences *D* as stated in Definition 2 and given that *C* is fine-grained (see below). I leave to the reader verification of my claims about Axioms 2–4. For Axiom 5, however, a formal observation is motivated, since this proof requires a few technical assumptions.

Observation 1. Suppose that *C* is fine-grained (for every $c_m \succ c'_m \in C$ there is some c''_m such that $c_m \succ c''_m \succ c'_m$) and also suppose that there is a set of bounded undesirable consequences $D(\subset C)$. Then utilitarianism, prioritarianism, and egalitarianism all entail Axiom 5 or Axiom 5′, and the maximin view entails a version of Axiom 5 according to which d_m and $c_{m'}$ have to be chosen such that the worst consequence of the act in question is not diminished.[11]

[10] For technical formulations of these ethical theories, see the proof of Observation 1.

[11] Expressed in technical terms, d_m and $c_{m'}$ should be chosen such that $C_m \cdot d_m \succcurlyeq C_{m'}$ and $C_{m'} \cdot c_{m'} \succcurlyeq C_m$.

The version of Axiom 5 entailed by the maximin view is trivially implied by the original version of Axiom 5. Furthermore, since this version is also sufficient for deriving the theorem stated in the next section—I leave it to the reader to verify this[12]—it can be reasonably maintained that the axiomatization constructed here is also relevant for advocates of the maximin view.

Thus, given that Axiom 1 can be accepted, or some other axioms that allow us to construct a cardinal utility function, it seems that we have a compelling argument in favor of utilitarianism, namely that this theory follows (see section III and the proofs) from premises that are accepted by advocates of all major positions in the debate about consequentialism.

III. TWO UTILITARIAN THEOREMS

Taken together, Axioms 1–5 imply that utility numbers can be assigned to all consequences, and that the solution to an ethical problem is the set of acts having the highest sum of utility.

Theorem 1. Let Axioms 1–5 hold for a set Δ of ethical problems. Then there is a real-valued function u on C such that for every $\delta \in \Delta$:

(1) $u(c \cdot c') = u(c) + u(c')$ and $u(c) > u(c') \Leftrightarrow c \succ c'$

(2) If u' is another function satisfying (1), then for some real number $r > 0$, $u = ru'$.

(3) $\Phi(\delta) = \{a : \sum_{i \in I} u(C_{ai}) \geq \sum_{i \in I} u(C_{a'i}) \text{ for all } a' \in A\}$

Parts (1) and (2) of Theorem 1 assert that the utility of consequences are measurable on a ratio scale. Part (3) is the utilitarian property, recommending us to choose alternatives that maximize utility.

The strongest axiom employed in Theorem 1 is Axiom 5, as noted in the previous section. While that axiom requires the value of a good consequence compensating for a poor consequence to be entirely determined by the poor consequence in question, the following weakening (Axiom 5′) allows the compensation to be a function of more factors, as long as it can be described by a positive monotone function.

Axiom 5′. (*Weak Trade-Off*) There is a set of bounded undesirable consequences D_{ai} such that for every $d_{ai} \in D_{ai}$ and every ethical problem δ, there is a (much better) consequence $c_{a'i'}$ determined by some positive monotone function (of α, d_{ai} and $c_{a'i'}$) such that, if $a \in \Phi(\delta)$ and $C_{ai} \succ C_{a'i'}$, it holds that $\Phi(\delta) = \Phi(\delta')$, where δ' is the ethical problem obtained from δ by substituting C_{ai} for $C_{ai} \cdot d_{ai}$ and $C_{a'i'}$ for $C_{a'i'} \cdot c_{a'i'}$.

[12] Note that the proof given below leaves the minimal level of act a''' in δ'' at least as high as the minimal level of act a' in δ''

Theorem 2, stated below, shows that also Axioms 1–4 and 5′ are sufficient for obtaining the desired support for utilitarianism.

Theorem 2. Let Axioms 1–4 and 5′ hold for a set Δ of ethical problems. Then there is a real-valued function u on C such that for every $\delta \in \Delta$, (1)–(3) in Theorem 1 are satisfied.

In my view, axiomatic arguments are valuable mainly because they show how certain central ideas are connected to each other. If a set of axioms is less controversial than some theorem, then a proof that the theorem follows from the axioms indicates that either one should give up at least one of the axioms, or accept the theorem. As I see it, the conjunction of Axioms 1–5, respectively Axioms 1–4 and 5′, are less controversial than the utilitarian doctrine. Therefore, it is an interesting piece of information that utilitarianism can be deduced out of those axioms. As pointed out in the introduction, the present axiomatization differs from Harsanyi's and his successors' axiomatizations in at least two aspects. First, Harsanyi constructed his axiomatization by adding a set of axioms to Bayesian decision theory. That is, his ethical conclusion is derived from what is fundamentally a set of rationality requirements for individual decision making under risk. Anyone who considers that theory of rationality to be unacceptable, for example, because he or she feels that the much criticized independence postulate (or "the sure-thing principle") lacks intuitive justification, will also refute Harsanyi's argument for utilitarianism.[13]

The second difference, which is perhaps less important, has to do with interpersonal utility comparisons. Harsanyi's Axiom *D* requires that agents can judge whether the utility functions for a set of n individuals are expressed in equal utility units.[14] This, he admitted, presupposes an ability to make cardinal comparisons of interpersonal utility. In the present axiomatization, we instead assume that ordinal comparisons of interpersonal utility are possible, and we show that a sufficiently rich structure of ordinal comparisons allows us to assign cardinal numbers (by applying Axiom 1). This is a well-known maneuver in measurement theory, and I do not want to claim that it is a radical improvement over Harsanyi's proposal. (For example, we still lack a satisfactory theory about what sort of evidence can legitimately be taken to indicate that one person enjoys a higher utility level than another.) A *small* advantage of the suggested approach, however, is that it does not, unlike Harsanyi's Axiom *D*, require direct numerical

[13] In Broome's version of Harsanyi's axiomatization, the most controversial Bayesian axiom is the impartiality axiom.

[14] Harsanyi, "Bayesian Decision Theory, Rule Utilitarianism, and Arrow's Impossibility Theory," p. 294.

comparisons; the numbers used for describing different utility levels are derived from the axiomatization itself.

MARTIN PETERSON

Royal Institute of Technology/Sweden

PROOFS

Proof of Observation 1. Utilitarianism: According to this theory, there is a function u that assigns real numbers to all elements in C, and for every ethical problem δ it holds that $\Phi(\delta) = \{a: u(C_{a1}) + ... + u(C_{an}) \geq u(C_{a'1}) + ... + u(C_{a'n})$ for all $a, a' \in A\}$. In order to derive Axiom 5, choose for every $d_{ai} \in D_{ai}$ a c'_{ai} such that $-u(d_{ai}) = u(c'_{ai})$. (That there is such a consequence c'_{ai} follows from the assumption that C is fine-grained.) Then the total sum of utility of the modified acts in δ and δ' will be equal; hence, these ethical problems have the same solution.

Prioritarianism: Prioritarians distinguish between utility and moral value; the moral value $v(a)$ of an act can be described by the equation $v(a) = r(u(C_{a1})) + ... + r(u(C_{an}))$, where r is a strictly increasing and concave function. In order to derive Axiom 5′ choose, for every $d_{ai} \in D_{ai}$ a c'_{ai} such that $r(u(C_{ai} \cdot d_{ai})) + r(u(C_{a'i} \cdot c'_{a'i})) = r(u(C_{ai})) + r(u(C_{a'i}))$. (That there is such a consequence $c'_{a'i}$ follows from the assumption that C is fine-grained.) Then the total sum of moral value of the modified acts in δ and δ' will be equal; hence, these ethical problems have the same solution.

Egalitarianism: According to this theory, the moral value $v(a)$ of an act can be described by the equation $v(a) = e(u(C_{a1}),..., u(C_{an}))$, where e is a function that is strictly increasing with respect to some appropriate measure of equality. In order to derive Axiom 5, choose, for every $d_{ai} \in D_{ai}$ a $c'_{a'i}$ such that $e(..., u(C_{ai}), u(C_{a'i}) ...) = e(..., u(C_{ai} \cdot d_{ai}), u(C_{a'i} \cdot c'_{a'i}),...)$. (That there is such a consequence follows from the assumption that C is fine-grained.) Then the total sum of moral value of the modified acts in δ and δ' will be equal; hence, these ethical problems have the same solution.

The Maximin View: This view does not presuppose any cardinal measurement of consequences. As noted in Observation 1, we consider a slightly weakened version of Axiom 5, according to which d_{ai} and $c_{a'i}$ are chosen such that $C_{ai} \cdot d_{ai} \succcurlyeq C_{a'i}$ and $C_{a'i} \cdot c_{a'i} \succcurlyeq C_{ai}$. In order to derive this version of Axiom 5 it is sufficient to note that the maximin view singles out the set of morally right acts by just taking into consideration the worst consequences of each act; hence, since the worst consequences of each act were assumed to be unaffected, and we know that there is a set of bounded undesirable consequences,

Calcutta University Library
CENTRAL CAMPUS

any $c_{m'}$ that satisfies the condition stated in Observation 1 can be chosen.

Proof of Theorem 1. Parts (1) and (2) rely entirely on Axiom 1, and are essentially an application of a standard theorem in measurement theory that has been carefully proved by others. Therefore, we will just provide a rough sketch of the main idea.[15]

In order to construct the function u we select a consequence c^* in C, which will serve as the unit of the ratio scale. For any other c in C and for any positive integer n, the Archimedean axiom 1:(iv) implies that there is an integer m such that $mc^* \succ nc$. Let m_n be the smallest integer for which this is true, and observe that $m_n c^* \succ nc \succcurlyeq (m_n - 1)c^*$. Hence, m_n copies of the unit consequence c^* are approximately equal to n copies of c. By selecting a larger and larger n the approximation becomes better and better, and, given that the limit exists, we can define u as: $u(c) = \lim_{n\to\infty} m_n/n$. In order to complete the proof, one has to show that the limit exists and that it satisfies the properties stated in (1) and (2).

Part (3) is easier to prove if we first prove the following lemma.

Lemma 3. (*Equal Trade-Off*) There is a closed interval of positive numbers $[a, b]$ such that for all $\alpha \in [a, b]$ and for all ethical problems δ, if $a' \in \Phi(\delta)$ and $u(C_m) > u(C_{m'})$, it holds that $\Phi(\delta) = \Phi(\delta')$, where δ' is the ethical problem obtained from δ by subtracting α from $u(C_m)$ and adding α to $u(C_{m'})$.

Because of parts (1) and (2) in Theorem 1 we know that utility numbers can be attached to all elements in C. Thus, by applying parts (1) and (2) to Axioms 3, 4, and 5, they can be restated in numerical terms. It is fruitful to state the numerical version of Axiom 5 as a trivial lemma. In Lemma 4 below, the concept of 'bounded undesirable consequences' in Axiom 5 has been replaced by a closed interval of real numbers, which property (1) allows us to do.

Lemma 4. There is a closed interval of positive numbers $[a, b]$ such that for all $\alpha \in [a, b]$ there is a number β such that for all ethical problems δ, if $a \in \Phi(\delta)$ and $u(C_m) > u(C_{m'})$, it holds that $\Phi(\delta) = \Phi(\delta')$, where δ' is the ethical problem obtained from δ by subtracting α from $u(C_m)$ and adding β to $u(C_{m'})$.

We are going to prove Lemma 3 by showing that $\alpha = \beta$ whenever

[15] For full details, see, for example, Krantz (*op. cit.*, pp. 77–81); see also F. Roberts, *Measurement Theory*, Volume 7 of Gian-Carlo Rota, ed., *Encyclopedia of Mathematics and Its Applications* (Boston: Addison-Wesley, 1979) For the sketch provided here, see Krantz, p. 75.

Lemma 4 is applied to an ethical problem. Assume for reductio that $\alpha \neq \beta$ as Lemma 4 is applied to the act a and the individuals i and i'. Let us first consider the case in which i and i' are the only individuals in δ, and $a \in \Phi(\delta)$. Let $u(C_{ai}) = u_1$ and let $u(C_{ai'}) = u_2$; $u_1 > u_2$. By applying Axiom 3 twice, we find that δ can be transformed into an ethical problem δ' with three new (identical) acts a', a'', a''' such that $\Phi(\delta') = \{\Phi(\delta) - a\} \cup \{a' \cup a'' \cup a'''\}$.

We now apply Lemma 4 to a'' and a''' in δ' exactly k times with some α' and β', such that $0 < u_1 - k\alpha' - [u_2 + k\beta'] < \alpha$. (It might hold that there are no k, α', β' that satisfy this inequality, for example, if $\beta' \gg \alpha'$. In that case, let $k = 1$, $\alpha' = \alpha$, $\beta' = \beta$, and ignore the next sentence.) Then we apply Lemma 4 to a''' one more time, this time subtracting α respectively, and adding β. It now holds that $u(C_{a''i}) < u(C_{a''i'})$ in δ''. Finally, we apply Lemma 4 again, but this time we subtract α from $u(C_{a''i'})$ and add β to $u(C_{a''i})$.

δ'			δ''		
	i	i'		i	i'
...			...		
a'	u_1	u_2	a'	u_1	u_2
a''	u_1	u_2	a''	$u_1 - k\alpha'$	$u_2 + k\beta'$
a'''	u_1	u_2	a'''	$u_1 - k\alpha' - \alpha + \beta$	$u_2 + k\beta' + \beta - \alpha$
...			...		

We initially assumed that $\alpha \neq \beta$. Hence, $-\alpha + \beta \neq 0$. Thus, by Axiom 4, $a'' \notin \Phi(\delta'')$ or $a''' \notin \Phi(\delta'')$, which contradicts the result that $\{a', a'', a'''\} = \Phi(\delta') = \Phi(\delta'')$. Consequently, our initial assumption must be false.

Let us now consider the case in which $a \notin \Phi(\delta)$. Then, by Axiom 2, there is some other act b such that $b \in \Phi(\delta)$. Furthermore, since β is a function of α, it follows that α and β can be applied to $u(C_{bi})$ and $u(C_{bi'})$ in the same way as above, yielding the analogous contradiction.

Finally, in order to handle the general case with more than two individuals, it is sufficient to note that for every individual $j \in \{I - \{i \cup i'\}\}$, it holds that $u(C_{aj}) = u(C_{ai'j}) = u(C_{a''j}) = u(C_{a''j})$. Hence, the maneuver from δ to δ'' can be carried out in a similar way as above. Q.E.D. (Lemma 3).

We are now in a position to prove part (3) of Theorem 1. Let δ be an arbitrary element in Δ. We then apply to δ the function to* (trade-off), which we define as the following function to iterated n times for some n such that $to^n(\delta) = to^{n+1}(\delta)$. Let *high*$(\delta, a)$ be a function that returns an ordered pair $\langle a, i \rangle$ such that $u(C_{ai}) \geq u(C_{ai'})$ for every i', and *low*(δ, a) a function that returns an ordered pair $\langle a, i \rangle$ such that

$u(C_{w})$ ≥ $u(C_{w'})$ for every i'. Let α be a small positive nonzero number obtained from Lemma 2 such that for every $u(C_{w})$ there is an even number t such that $\alpha \cdot t = u(C_{w})$. (In case there are no α and t satisfying this condition, we also have to choose some additional number $\alpha' < \alpha$, that we apply the "last time" utility is subtracted; that is, α' should be chosen such that $\alpha \cdot t + \alpha' = u(C_{w})$.)

$$\text{to}(\delta) = \begin{cases} \delta \text{ if } u(C_{w}) = u(C_{w'}) \text{ for all } a, i, i' \\ \langle A', I', u' \rangle \text{ otherwise, with:} \end{cases}$$

(i) $A' = A$
(ii) $I' = I$
(iii) For some a such that $h = high(\delta, a)$, $l = low(\delta, a)$ and $u(h) \neq u(l)$, for all r, t, v, w:
$u'(c_{w}) = u(c_{w}) - \alpha$ if and only if $\langle a, i \rangle = h$
$u'(c_{w'}) = u(c_{w'}) + \alpha$ if and only if $\langle a, i' \rangle = l$
$u'(c_{w',v}) = u(c_{w',v})$ for all $\langle a', i'' \rangle \neq h, l$

Since rule to* is obtained from an iterated application of Lemma 2 (equal trade-off), it holds that $\Phi(\delta) = \Phi(\text{to}^*(\delta))$. It follows from the definition of α above that to* will converge such that $u(C_{w}) = u(C_{w'})$ for all a, i, i' in the ethical problem $\text{to}^*(\delta)$.

We now apply Axiom 4 and Axiom 2 to $\Phi(\text{to}^*(\delta))$, which gives us a set $A^* = \{a: \sum_{i \in I} u(C_{w}) \geq \sum_{i \in I} u(C_{w',i})$ for all $a' \in A\}$. Hence, $\Phi(\delta) = \Phi(\text{to}^*(\delta)) = A^* = \{a: \sum_{i \in I} u(C_{w}) \geq \sum_{i \in I} u(C_{w',i})$ for all $a' \in A\}$. Q.E.D. (Theorem 1).

Proof of Theorem 2. We prove Theorem 2 by showing that Theorem 1 holds even if Axiom 5 is substituted by Axiom 5′, that is, that property (3) of Theorem 1 follows from Axioms 1–4 and 5′. As before, we apply property (1) and (2) of Theorem 1 and obtain the following numerical version of Axiom 5′:

Lemma 5. There is a closed interval of positive numbers $[a, b]$ such that for all $\alpha \in [a, b]$ and for all ethical problems δ, there is a (much larger) number β determined by some monotone function (of α and u_{w} $u_{w'}$) such that if $a \in \Phi(\delta)$ and $u(C_{w}) > u(C_{w'})$, it holds that $\Phi(\delta) = \Phi(\delta')$, where δ′ is the ethical problem obtained from δ by subtracting α from $u(C_{w})$ and adding β to $u(C_{w'})$.

It will be sufficient to prove Lemma 3, since the proof of property (3) of Theorem 1 only relies on that lemma in conjunction with Axiom 4. Lemma 3 is proved by showing that $\alpha = \beta$ whenever Lemma 5 is applied to an ethical problem. Assume for reductio that $\alpha \neq \beta$ as Lemma 5 is applied to the act a and the individuals i and i'. Let us first consider the case in which i and i' are the only individuals in

δ, and $a \in \Phi(\delta)$. Let $u(C_m) = u_1$ and let $u(C_{m'}) = u_2$; $u_1 > u_2$. By applying Axiom 3 twice, we find that δ can be transformed into an ethical problem δ' with three new (identical) acts a', a'', a''' such that

$$\Phi(\delta') = \{\Phi(\delta) - a\} \cup \{a' \cup a'' \cup a'''\}.$$

We now apply Lemma 5 to a'' and a''' in δ' exactly k times with some α' and β', β'',..., such that $0 < u_1 - k\alpha' - [u_2 + (\beta' + \beta'' + ...)] < \alpha$. The difference between u_1 and u_2 is now very small. (If, for example, $\beta' \gg \alpha'$ it might hold that there are no k, α', β', β'' that satisfy this inequality; in that case, let $k = 1$, $\alpha' = \alpha$, $\beta' = \beta$, and ignore the next sentence.) Then we apply Lemma 5 to a''' one more time, this time subtracting α respectively adding β. It now holds that $u_{m-1} < u_{m-1'}$ in δ.

δ'

	i	i'
...		
a'	u_1	u_2
a''	$u_1 - k\alpha'$	$u_2 + (\beta' + \beta'' + ...)$
a'''	$u_1 - k\alpha' - \alpha$	$u_2 + (\beta' + \beta'' + ...) + \beta$
...		

δ''

	i	i'
...		
a'	u_1	u_2
a''	$u_1 - k\alpha'$	$u_2 + (\beta' + \beta'' + ...)$
a'''	$u_1 - k\alpha' - \alpha + \beta_0$	$u_2 + (\beta' + \beta'' + ...) + \beta - \alpha$
...		

We now apply Lemma 5 to a''' again and subtract α from $u_{m-1'}$ and add β_0 to u_{m-r}. This yields the ethical problem δ''. Since $\beta \neq \alpha$ there are two cases to consider, namely, $\alpha > \beta$ and $\beta > \alpha$. Let us first consider the case when $\beta > \alpha$. Then, $\beta_0 > \beta > \alpha$, because of monotonicity. Hence, $-\alpha + \beta_0 > -\alpha + \beta > 0$. Thus, by Axiom 4, we have $a'' \notin \Phi(\delta'')$ or $a''' \notin \Phi(\delta'')$, which contradicts the result that $\{a', a'', a'''\}] = \Phi(\delta') = \Phi(\delta'')$. Consequently, our initial assumption must be false. The case when $\alpha > \beta$ is handled in the analogous way.

The case in which $a \notin \Phi(\delta)$ is handled exactly the same as in Theorem 1, as is the general case with more than two individuals. *Q.E.D.* (Theorem 2).

COMMENTS AND CRITICISM

THE FIRST PERSON: ERROR THROUGH MISIDENTIFICATION, THE SPLIT BETWEEN SPEAKER'S AND SEMANTIC REFERENCE, AND THE REAL GUARANTEE*

In this article, I wish to contrast three different phenomena concerning the first person, which, so far, have not been sufficiently distinguished in the philosophical literature. A lack of clarity about their distinction has led philosophers such as Elisabeth Anscombe,[1] and, more recently, Carol Rovane[2] and Andrea Christofidou[3] to claim that all, or at least most uses of 'I' and its cognates are *immune to error through misidentification*. I wish to show how such a conclusion is mistaken and possibly brought about, at least in part, by the conflation between error through misidentification and what I will call the *split between speaker's reference and semantic reference*. Further reflection on the latter phenomenon, in turn, will show that while not all uses of 'I' are immune to that split, it is true that they all have a further much more fundamental and important guarantee, so far I believe unremarked, which I will call the *real guarantee*. It is this real guarantee, I suggest, that Anscombe, Rovane, and Christofidou are in fact trying to account for under the misapprehension that it is what others, such as Sydney Shoemaker, Gareth Evans, Crispin Wright, and John McDowell, have called immunity to error through misidentification.[4]

I. ROVANE'S AND CHRISTOFIDOU'S ACCOUNTS OF ERROR THROUGH MISIDENTIFICATION

As is familiar, the idea of immunity to error through misidentification can be traced back to Ludwig Wittgenstein's distinction, in *The Blue*

* I would like to thank the Fulbright program and the Fonds National Suisse EFRIFO406101443 for their support.

[1] "The First Person," in Samuel Guttenplan, ed., *Mind and Language. Wolfson College Lectures 1974* (New York: Oxford, 1975), pp. 45–64; reprinted in Palle Yourgrau, ed., *Demonstratives* (New York: Oxford, 1990), pp. 135–58 (page reference to the latter)

[2] "The Epistemology of First-person Reference," this JOURNAL, LXXXIV, 3 (March 1987): 147–67; "Self-reference: The Radicalization of Locke," this JOURNAL, XC, 2 (February 1993): 73–97—here see p. 90; and *The Bounds of Agency* (Princeton: University Press, 1998).

[3] "First Person. The Demand for Identification-free Self-reference," this JOURNAL, XCII, 4 (April 1995): 223–34.

[4] Shoemaker, "Self-reference and Self-awareness," this JOURNAL, LXV, 19 (October 3, 1968): 555–67; Evans, *The Varieties of Reference* (New York: Oxford, 1982); Wright, "Self-Knowledge: The Wittgensteinian Legacy," in Wright, Barry C. Smith, and Cynthia McDonald, eds., *Knowing Our Own Minds* (New York: Oxford, 1998), see pp. 13–45, McDowell, "Reductionism and the First Person," in Jonathan Dancy, ed.,

0022-362X/03/0008/416–431 © 2003 The Journal of Philosophy, Inc.

Book, between what he called uses of 'I' (or 'me' and 'my') *as subject* and its uses *as object*.[5] Examples of the former kind of use are "I have a toothache," "I see so-and-so"; examples of the latter, by contrast, are "My arm is broken," "I have a bump on my forehead." The difference between the two kinds of use, according to Wittgenstein is this: in the latter case I could utter, "My arm is broken," say after a car accident, on the basis of feeling pain in my arm and of seeing a broken arm at my side. Yet, that self-ascription could be based on having mistaken the broken arm at my side for mine, when, in fact, it is someone else's.[6] In such an event, there would be someone's arm that is broken but I would be mistaken in thinking that it is mine. By contrast, in the former case, there seems to be no provision for this kind of error. I may be wrong about what I am seeing, for instance, but there is no issue but that it is I who am doing the seeing.[7]

Here, however, is how Rovane characterizes error through misidentification:

> Let me describe, therefore, a situation in which we *might* [italics added] be prepared to grant that a use of the first person is accompanied by error through misidentification. Suppose I am facing a mirror and I believe that I see my own reflection when I really see someone else's. On the basis of what I see reflected in the mirror I say, "There's an incredibly tasteless painting hanging on the wall directly behind me." Because I believe that I am the person reflected in the mirror, *I take 'me' to refer to the person reflected in the mirror* [italics added]. Thus there is a sense in which I take someone who is not myself, namely, the person reflected in the mirror, to have made my utterance and, hence, to be the referent of that use of 'me'. *As is usual with cases of error through misidentification, we can interpret the utterance so as to make it come out true or false, depending on what referent we prefer to assign to the offending expression* [italics added].... We are inclined to say that I must have intended to

---

Reading Parfit (Malden, MA: Blackwell, 1997), pp 230–50. See also Christopher Peacocke, *Sense and Content* (New York: Oxford, 1983); and J. Pryor, "Immunity to Error through Misidentification," *Philosophical Topics*, XXVI, 1 (1998): 271–304.

[5] Cf *The Blue and the Brown Books* (Oxford: Blackwell, 1958), see pp 66–67.

[6] In the course of this article, I will mostly talk about self-ascriptions that are liable to error through misidentification, which may or may not be linguistically expressed I will also use the following conventions: single or double inverted commas for words and sentences, respectively, and square brackets for concepts and propositional contents.

[7] Nowadays there is a widespread agreement on the fact that immunity to error through misidentification is *not* a phenomenon that is characteristic only of self-ascriptions and that, for instance, uses of demonstratives can also be immune to that form of error. There is also widespread consensus about the fact that, contrary to what seems likely to have been Wittgenstein's position on the matter, the distinction between error and immunity to error through misidentification, relative to the first person, cuts across the physical and the psychological domains.

> refer to myself when I used the expression 'me' and that I thought that *in* referring to myself I was also referring to the person reflected in the mirror. Hence it is more natural to interpret my utterance as false of myself rather than as true about the person reflected in the mirror. Nevertheless, *the alternative interpretation might be warranted in certain circumstances, and, when it is, we have a use of the first person that incorporates error through misidentification* [italics added]. We must grant, then that, although uses of 'I' can be subject to referential error through misidentification, it is extremely rare that they are. There is always the strongest presumption that a speaker intends to speak of *herself* in using 'I'.[8]

I have quoted at length because the passage clearly indicates that, on the basis of the example she discusses, Rovane thinks the following:

(1) that if error through misidentification relative to the first person occurs, then 'I' and its cognates are not used to refer to the subject but to the person who is somehow (in this example, visually) presented to her;
(2) that—for reasons we will see in a moment—in the case of the use of the first person, it is merely *extremely rare* that it is used to refer to someone else. As a consequence, Rovane holds
(3) that *most* uses of the first person are immune to error through misidentification.

Before addressing assumption (1), which is the leading thought, let me briefly present and discuss Rovane's various defenses of (2).

Rovane thinks that reference to objects, either in speech or in thought, is always mediated by some belief(s) about the referents, which would determine which object one is speaking or thinking about. Her account of which belief(s) should be in place in order to secure self-reference, however, varies across her writings. On the one hand, it is held that "speakers generally have enough true beliefs about themselves in virtue of which they are quite clear about their identities even though they have some false beliefs about themselves as well."[9] Hence, it is because it is very seldom the case that one has *massively* false beliefs about oneself that there is always the strongest presumption that in using the first person one is thinking or talking about oneself. On the other hand, since, on this view, it could hardly be explained how amnesiac and seriously deluded subjects could manage to refer to themselves by using 'I', Rovane presents two other identity beliefs to mediate first-person reference, namely [I = the

[8] "The Epistemology of First-person Reference," pp. 153–54. Rovane introduces error through misidentification essentially in the same way in her more recent work—cf. *The Bounds of Agency*, pp. 220–21.

[9] "The Epistemology of First-person Reference," p. 154.

person who is speaking now] and [I = the thinker of this thought].[10] But these beliefs seem to be a priori true, for, whenever I speak, it is true that I am the person who is speaking now. Moreover, if it is a priori true that I can only be introspectively aware of my own thoughts,[11] then the identity between *I* and *the thinker of this thought* would hold a priori. Yet, on both these views, assuming further that (1) is correct, *all* uses of the first person would be immune to error through misidentification, contrary to what Rovane maintains—that error through misidentification *can*, exceptionally, afflict uses of the first-person pronoun and its cognates.

This is in effect Christofidou's position, although in her article she reacts to Rovane's account on the grounds that the *reliability* of [I = the thinker of this thought] would never secure *immunity* to error through misidentification, but only, at most, that one's self-ascriptions—which remain potentially open to that kind of error—are in fact not affected by it (*op. cit.*, p. 226). Here are two relevant passages:

> Even if my memories and perceptions deceive me, and it is not *I* who climbed the steps of St. Paul's but a friend, then I misidentify, not myself, but the person who climbed the steps; my use of 'I' refers correctly to me, and refers to me directly or unmediatedly. In discovering that I did not after all climb the steps of St. Paul's, then I should say 'so I did not climb them', and then I should try to explain how I could have fallen into error. But the error has nothing to do with the self-reference of 'I'. If there is a problem with propositions such as 'I climbed the steps of St. Paul's' the mistake is in the predication component, *not in the identification component*. I have *misascribed* to myself a predicate, but the reference or identity of 'I' is unshaken: *'I' is immune to error through misidentification, whatever the predicate might be* (*op. cit.*, p. 227; italics added).

> [D]iscussions in this area have been vitiated by the failure to distinguish between two types of immunity: immunity to error through misidentification (which covers *all* uses of 'I' whether these involve corporeal predicates, or mental predicates, or both) and what I call immunity to error through *misascription* (which covers only certain mental predicates) (*op. cit.*, footnote 7).

Here again, however, the startling claim

(3′) that *all* uses of 'I' are immune to error through misidentification

is reached by holding

[10] The first proposal is made in "The Epistemology of First-person Reference," and in *The Bounds of Agency*, pp. 221–22; the second in "Self-reference."

[11] I defend this claim in my "Thought Insertion and Immunity to Error through Misidentification," *Philosophy, Psychology and Psychiatry*, IX, 1 (2002): 27–34.

(1) that if error through misidentification occurs, then 'I' would not refer to the subject but to someone else and
(2′) that this is impossible, because 'I' refers directly and unmediatedly to the subject.[12]

Now, claim (2′) allows of two interpretations. It can be taken to mean that 'I' is an *automatic indexical*, which invariably refers to the producer of its tokens.[13] Hence, all tokens of 'I' refer to the person who produced them. Or else, it can be taken to mean that since linguistic self-reference, in a framework like the one elaborated by Evans in *The Varieties of Reference* (*op. cit.*), for instance, is taken to be dependent on the ability to have I-thoughts and these, in turn are grounded on *identification-free* ways of having knowledge about oneself, then *all* self-reference by means of the first person is direct and unmediated.[14] I surmise that Christofidou has this second reading of (2′) in mind. Yet, it is striking—and indeed it should have struck Christofidou—that Evans for one *never* claimed that *all* uses of 'I' were immune to error through misidentification, while, for the reasons just given, he indeed thought that linguistic *self-reference* was unmediated.

Before turning to a discussion of assumption (1), let me point out an important difference between Christofidou's treatment of the relevant class of self-ascriptions and Rovane's. Christofidou's view—contrary to Rovane's—allows them to be false, because, on her view, 'I' would always refer to the subject who, however, may not instantiate the property expressed by the predicate. Yet she thinks that that would be just a common mistake of *misascription*—as opposed to a mistake of misidentification—relevantly similar to the case in which, for instance, being deluded about the color of the shirt I am wearing, I falsely say, "I am wearing a white shirt."

Suppose, however, that my claim "I am wearing a white shirt" was

[12] In fact, the particular example given by Christofidou is problematic. For on some interpretations (cf McDowell) it would turn out to be immune to error through misidentification, while on others (cf. Pryor), it would not. It all depends on how we choose to interpret cases of quasi-memories and on what account we give of Shoemaker's distinction between de facto and logical immunity to error through misidentification, if we accept it at all. This, however, is not an issue that I can pursue within the scope of this article.

[13] Cf. John Perry, "Indexicals and Demonstratives," in Bob Hale and Wright, *A Companion to the Philosophy of Language* (Malden, MA: Blackwell, 1997), pp. 595–98 for the use of the label "automatic indexical" and its discussion.

[14] Knowledge of the truth of a thought such as [I am *F*] is identification-free, according to Evans, if it is not based on knowledge of the truth of a pair of thoughts such as [*a* is *F*] and [I = *a*], but is just based on receiving information directly about oneself, either through perception and proprioception, or through introspective awareness of one's own mental properties, or by having stored information about oneself in memory—cf. Evans, chapter 7.

made on the basis of seeing someone reflected in a mirror wearing a white shirt and of wrongly taking that person to be me. Suppose, further, that, by chance, it was really the case that I was wearing a white shirt. Now, on Christofidou's account of the matter, since we cannot say that my use of 'I' is affected by error through misidentification, we are obliged to say that my claim is perfectly in order, for, in this case, there is no misascription of a predicate, after all. But this seems to be highly counterintuitive. Again, this counterintuitive view would depend on having excluded the possibility that uses of 'I' can be affected by error through misidentification, on the ground that if they were, they would not refer to the subject.

To sum up: despite their differences, Rovane's and Christofidou's accounts of error through misidentification and of the corresponding kind of immunity are based on a structurally similar piece of reasoning, which assumes that if error through misidentification occurs, then 'I' and its cognates would be used to refer to someone other than oneself. Recoiling from this idea, they try to qualify the conclusion by, respectively: attempting a semantic explanation (Rovane), appealing to the extraordinary reliability (which is in fact the a priori truth) of the identity belief that allegedly mediates first-person reference, namely [I = the thinker of this thought]; and by denying (Christofidou) that any belief mediates first-person reference at all.

Both their accounts seem problematic, however, for a striking consequence of their views is that, for *conceptual* reasons, error through misidentification is never or hardly ever possible. This runs contrary to what has been variously maintained in the rest of the literature on the topic,[15] and to some solid intuitions. For, after all, it seems quite intuitive to say that, in Rovane's own example, for instance, the claim "There is an incredibly tasteless painting hanging on the wall behind me" would be affected by error through misidentification, relative to the first person, if the person reflected in the mirror were in fact someone else. In such an event, my claim would not be false because the person reflected in the mirror does not instantiate the property ascribed to her by the predicate. It would be false *because I am not that person.*

II. ERROR THROUGH MISIDENTIFICATION: THE CORRECT ACCOUNT

Let us now turn to assumption (1), on which both Rovane's and Christofidou's accounts of self-ascriptions that are immune to error through misidentification ultimately rest: the assumption that if error through misidentification occurred (or could occur), then one would

[15] Cf. footnote 4.

not be saying something about oneself, but about the person one has mistakenly taken to be oneself, so that 'I' (or its cognates) would refer not to its utterer, but to someone else.

In order to appraise this assumption, let me propose a template. Consider Rovane's example again. Here is how the subject's mistake can be accounted for. The self-ascription "There is an incredibly tasteless painting hanging on the wall behind me" can be seen as rationally *grounded* on the following pair of beliefs: [There is an incredibly tasteless painting hanging on the wall behind the person reflected in the mirror], and [I am the person reflected in the mirror]. Call these beliefs the *predication* and the *identification component*, respectively. These beliefs rationally sustain the subject's final self-ascription in the following sense: while she may not consciously entertain them in order to move to her self-ascription, still we would expect her to appeal to them in order to justify it, if challenged, and to withdraw from it if one of the two sustaining beliefs turned out to be false.[16]

In Rovane's view, as we have seen, the resulting self-ascription may be regarded as true or false depending on whether we take the use of 'I'/'me' to refer either to the person actually reflected in the mirror, or to the subject. Rovane also insists, however, that, although it is more natural to take the final self-ascription as saying something false of the subject, "the alternative interpretation might be warranted in certain circumstances, and, when it is, we have a use of the first person that incorporates error through misidentification."[17] This is odd, however. For, if this claim is true, what does the error consist in? There seem to be three possibilities.

According to a first reading, we might take Rovane to be saying the following: the final self-ascription is affected by error through misidentification because it is grounded on a false identification component; yet it is true because 'I' should be taken to refer to the person who is actually reflected in the mirror. This, however, would require that the first person be used ambiguously across the inference. For the identification component is wrong only if the occurrence of the first person in it refers to the person standing in front of the mirror, rather than to the person she sees reflected in the mirror. The cost of this interpretation is accordingly that the resulting self-ascription would be unwarranted, because it would be based on an equivocation between the occurrence of the first person in the premise and its occurrence in the conclusion. More explicitly, on this first reading

[16] For a discussion of this issue, cf. Peacocke, pp. 143–44

[17] "The Epistemology of First-person Reference," pp. 153–54.

we can represent the inferential grounds, which rationally sustain the self-ascription, as follows:[18]

(a) (I) A is F
(II) $A = I_B$
(III) I_A am F

where, unbeknownst to the subject, $A \neq I_B$. However, (a) is not valid, since the use of the first person in the second premise and in the conclusion would refer to two different people. So, on this reading, any case of error through misidentification would be a case of *faulty reasoning leading to a true belief*.[19]

According to the second possible reading, we interpret the use of the first person across the inference as uniformly referring to the person visually presented to the subject. As a consequence, however, we would have that the identification component on which the final self-ascription is based is, surprisingly, correct. For if the first person refers to the person who is actually reflected in the mirror, then surely she is identical to herself. Yet, on this reading, we would have the startling conclusion that error through misidentification occurs if a self-ascription is based on a *true* identification component![20] The form of inference that would rationally sustain the self-ascription would be as follows:

(b) (I) A is F
(II) $A = I_A$
(III) I_A am F

(b) would thus be valid and (III) true. But, then, where is the error?

[18] Where 'A' refers to the person actually reflected in the mirror, and 'B' refers to the subject who is making the self-ascription, and the subscripts indicate who is actually referred to by an occurrence of the first person.

[19] In conversation, Rovane has distanced herself from this reading and has proposed the following:

(I) A is F
(II) A = The person I take to be me
(III) The person I take to be me = I_B
(IV) I_B am F

This inference is valid alright, but the reader will be able to verify that it would in fact be a special case of (c) below and that the conclusion would be *false*, contrary to what Rovane claims should happen when error through misidentification occurs.

[20] The same kind of arguments would apply against Christofidou. For, *if* error through misidentification *could* occur, then the final self-ascription would not be about myself, but about the person reflected in the mirror. But this could only be the case if either the self-ascription is based on an equivocation between the occurrence of the first person in the premise and in the conclusion, or else, if it is based on a correct identification component, which is absurd.

There is only one last possibility: to adopt a uniform reading of the first person across the inference, yet, this time, as referring to the subject who is making the self-ascription. The result would be that the final self-ascription is false and it is so because it is based on the false identification component [I = the person reflected in the mirror]. The form of inference, which would rationally sustain the final self-ascription, can then be represented as follows:

(c) (I) A is F
(II) $A = I_s$
(III) I_s am F

(c) is valid. Yet, unbeknownst to the subject, $A \neq I_s$. Hence, the conclusion is false.

This, I submit, is the right account of the matter. Yet, to require that, in order to account for error through misidentification, the first person be interpreted across the inference as uniformly referring to oneself, is actually *inconsistent* with both Rovane's and Christofidou's assumption (1) according to which, if error through misidentification occurs, then the use(s) of the first person would refer to someone else.

A satisfactory account of error through misidentification relative to the first person should then respect the following features:

(1) error through misidentification can occur, although it may be contingently rare;
(2) when it does, it may issue in a false belief, correctly expressed;
(3) yet that belief is rationally grounded and this requires
(4) continuity of reference to oneself throughout the belief and its grounds.

Returning to Rovane's example, this means that "There is an incredibly tasteless painting hanging on the wall behind me" is affected by error through misidentification precisely because it is a statement about *me*—and not about the person reflected in the mirror whoever she is. Therefore, that self-ascription is false because it is not true of *me* that I instantiate the property expressed by the predicate and I make it because I wrongly take the person reflected in the mirror to be me.

III. THE DIAGNOSIS I: THE SPLIT BETWEEN SPEAKER'S AND SEMANTIC REFERENCE

The question we should now ask is why one may be tempted to hold Rovane and Christofidou's assumption (1). My suggestion is that this may be at least partly due to a conflation between error through misidentification and the second phenomenon I am here concerned

with: the (im)possibility of a split between speaker's and semantic reference.

Anscombe, in her famous paper on the first person, argues that 'I' has certain guarantees, which other singular expressions do not have. In particular, she claims that uses of 'I' are guaranteed of having a referent, because the very utterance of a token of 'I' presupposes and, therefore, guarantees its existence.[21] She also claims that any given use of 'I' has a further guarantee as well: that the person that the user takes to be its referent is its referent. She phrases her point as follows:

> Guaranteed reference for...['I']...would entail a guarantee, not just that there is such a thing as *X*, but also that what I take to be *X* is *X* (*op. cit.*, pp. 145–46).

Anscombe calls this latter guarantee "Immunity to error through misidentification." She goes on to characterize it as follows:

> An 'I'-user cannot take the wrong object to be the object he means by 'I'. (The bishop may take the lady's knee for his, but could he take the lady herself to be himself?) (*ibid.*).

Anscombe's thought could appear to be this, that for any given use of 'I' to be comprehending, it must be such that one uses it intending to refer to oneself. Since, according to Anscombe, one cannot take someone else to be oneself, any use of 'I' is guaranteed of reference to its proper semantic referent, namely oneself.

Contrast 'I', so viewed, with a proper name like 'Dummett'. For a token use of the latter to be comprehending it must be made intending to refer to Dummett, but since I can take the wrong person to be Dummett, my use of that name is not guaranteed of reference to its actual semantic referent. The idea is that one can use a proper name comprehendingly, that is, intending to refer to its semantic referent, while not being able to answer the question, "Which object is *a*?" for instance, by saying (correctly), "That (perceptually presented) person is *a*." This, in turn, opens up the possibility that one may use 'Dummett' intending to refer to Dummett, but, in effect, not knowing which person is Dummett, take the wrong person to be him and use that name to refer to that person. In short: failure of knowing which person is the semantic referent of a singular expression allows one to use that expression to refer to someone other than its semantic

[21] Evans notoriously contests this guarantee, but this is not the place to discuss this issue.

referent, if the latter is mistakenly identified with the wrong object. Call this phenomenon *the split between speaker's and semantic reference*. Such a split was made familiar by Keith Donnellan,[22] although his classic discussion restricted it to the use of definite descriptions.

Anscombe's use of "immunity to error through misidentification" thus applies to something we might better call "impossibility of a split between speaker's and semantic reference." For it consists in the (putative) guarantee that if one's use of 'I' is competent, one must intend to use it to refer to oneself and that, in effect, the object one takes to be oneself is indeed oneself. This guarantees that the proper semantic referent of 'I'—namely, oneself—and the object that one intends 'I' to refer to, on a given occasion—that is, the "speaker's referent" of 'I'—be one and the same.

Now, a conflation between error through misidentification and a split between speaker's and semantic reference may well provide at least part of the motivation for the claim that if error through misidentification could arise, then one's use of 'I' would not refer to oneself but to the person who is wrongly taken by the subject to be oneself. Recall, however, that Anscombe (as well as Rovane and Christofidou) holds that this may never (or almost never) occur, in the case of the use of 'I' *on conceptual grounds*.

On reflection, however, it is obviously not true that all uses of 'I' are immune to a split between speaker's and semantic reference. For, while intending to use 'I' to refer to myself, I can wrongly take someone else for me, on a given occasion, and, hence, use 'I' to speaker-refer to that person. The usual kinds of example suffice to note this point. Consider, for instance, the case of a person who uses 'me' to refer to someone, different from herself, she sees reflected in the mirror. If there is a sense in which she can take that person to be herself, then there is a sense in which her use of 'I' can refer to that person. I think there is such a sense. For suppose that someone had just said, "There are no tasteless paintings in this room," and that, seeing the painting seemingly hanging behind her, the speaker feels in a position to contradict her interlocutor, by pointing out to him that, after all, there is one incredibly tasteless painting hanging on a wall of the room. In that case, if he then pointed out to her that the

[22] "Reference and Definite Descriptions," *Philosophical Review*, LXXV, 3 (1966): 281–304. Although Rovane is skeptical about the distinction between speaker's and semantic reference, she introduces error through misidentification in connection with Donnellan-like examples, which, in my view are best accounted for as cases of such a split—cf. "The Epistemology of First-person Reference," p. 152, and *The Bounds of Agency*, p. 220.

person behind whom there is a tasteless painting is not her, she would reply, "Well, there is an incredibly tasteless painting hanging on the wall behind *that* person reflected in the mirror." What her reply would show is that her original referential *intentions* in using 'I' were *primarily* directed towards that person, whom she mistakenly took to be herself. So, it is not true that all uses of 'I' and its cognates are immune to a split between speaker's and semantic reference. Once again, there does not seem to be any conceptual barrier to the possibility that such a split, however contingently rare it may be, can afflict the use of 'I'. Hence, to sum up:

(i′) a split between speaker's and semantic reference can occur in the case of the use of 'I', although it may be contingently rare;
(ii′) when it does, it may result in a true statement which
(iii′) is based on an ungrounded inference, and
(iv′) this, in fact, requires that the occurrence of the first person in the conclusion be used (in the kind of example I am considering here) to speaker-refer to an object other than oneself, somehow presented to one.[23]

So if what Anscombe, Rovane, and Christofidou are trying to highlight is some kind of sure-fire guarantee for our uses of 'I' and 'I' is immune neither to error through misidentification, nor to the split between speaker's and semantic reference, then we still have to find out what it is.

IV. THE DIAGNOSIS II: THE REAL GUARANTEE

In order to get clear about what Anscombe, Rovane, and Christofidou may be after, I think it is crucial to realize that there is an important difference between the case of a split between speaker's and semantic reference, which may occasionally afflict uses of 'I', and the case where that split afflicts uses of other singular expressions—proper names and definite descriptions in particular.[24] The difference is this. In the latter case (usually, although not necessarily)[25]

[23] Notice how the pattern of inference (a) presented in the previous section would fit with this account of the split between speaker's and semantic reference.

[24] Even someone who was skeptical about the possibility of a split between speaker's and semantic reference in the case of 'I' should acknowledge that *if* it could occur, then it would be structurally different from the same kind of split arising for other singular expressions.

[25] This, however, does not make any difference for our present purposes because, for the contrast to stand, it is sufficient that this *may* be the explanation of the split, at least in some cases of uses of proper names and definite descriptions, while it can *never* be the explanation of the split in the case of the use of 'I' (provided, as we will see, that 'I' is used comprehendingly).

(a′) the subject does not know which person is the *semantic* referent of the singular term;
(b′) this ignorance explains why she may take another person, who is different from the semantic referent of the singular term, to be the referent of that expression.

By contrast, in the former case, although a split can occur, that does not threaten the *presumption* that

(a″) the subject indeed knows which person she is.

Hence, ignorance of which person is the *semantic* referent of 'I' cannot be appealed to in order to explain why a subject may take someone else to be the person to whom a given use of 'I' by her refers. So, what explains it?

The correct form of explanation can be illustrated as follows:

(a*) the subject (to continue with Rovane's example) is ignorant of the identity *of the person visually presented to her*, and
(b*) she mistakenly identifies *that person* as herself.

The point is simply that one may well be able visually to discriminate a person—*that person*—yet still not be in a position to answer the question, "Which person is that one?" or, equivalently, "Who is that person?" Then, having wrongly identified that person as oneself, one goes on to use 'I' to speaker-refer to that person. Yet, as Christofidou would put it, one has not misidentified oneself, but *that person*. That is to say, although one may use 'I' in order to purport reference to that person because one does not know who that person is and wrongly takes her to be identical to oneself, one is under no misapprehension about *one's own* identity in doing so.

The guarantee possessed by competent uses of 'I' is then, not immunity to error through misidentification, nor to a split between speaker's and semantic reference, but this:

The Real Guarantee (at the level of language): the comprehending use of 'I' guarantees that the speaker knows which person is its *semantic* referent.

The real guarantee holds on purely conceptual grounds. It is a rule for the competent use of 'I' that one must use it to refer to *oneself*. But one cannot so much as be in a position to understand this rule unless one knows which person one is. This, in turn, implies that *qua* competent 'I'-user, the subject knows which person is the semantic referent of 'I'.

Notice, however, that the real guarantee must also operate at the level of thought. Hence, even if one were skeptical about the distinc-

tion between speaker's and semantic reference in the case of 'I', one could still retain the following:

> *The Real Guarantee* (at the level of thought): the possession of the first-person concept guarantees that the subject knows which person that concept is a concept of.

The real guarantee holds at the level of thought because the first-person concept is the concept of *oneself* and one cannot have it unless one knows which person one is. Hence, there is a guarantee that *qua* possessor of the first-person concept, the subject knows which person that concept is a concept of.

My claim then is that the phenomenon which Anscombe and, more recently, Rovane and Christofidou have really been concerned with, is not immunity to error through misidentification, nor immunity to a split between speaker's and semantic reference, but in fact the real guarantee. Indeed, a profitable way of interpreting the debate between Rovane and Christofidou is to see it as concerning how best to characterize one's knowledge of which person one is, which, in its turn, is supposed to ground one's grasp of the first-person concept both in language and in thought: while Rovane thinks of this knowledge as a kind of propositional knowledge of the form "I am *F*," or "I am the *F*," Christofidou thinks of it as some kind of nonpropositional, direct acquaintance with oneself.

Looking at their debate from this perspective can in fact help us to arbitrate it. For, if knowledge of oneself depended on having a set of (mostly) true beliefs about oneself, then it would follow that global amnesiacs would not know which persons they are. Hence, they could neither use 'I' comprehendingly, nor have thoughts about themselves. But, surely, we would like to be able to say that if a global amnesiac said (or thought), "I cannot remember anything about my past," her use of the first person would be perfectly in order. This, in turn, shows that knowing which person one is is *not constitutively* dependent on having such a set of true beliefs about oneself.[25]

[25] Indeed, if this proposal were meant to imply that, had one a set of false beliefs about oneself, then one might in fact be using the first person to think and talk about some other person of whom those beliefs are true, it would be doomed to failure for the following reason. In order for beliefs expressible as "I am *F*," or "I am the *F*" to be false, they should indeed be about *oneself*. Otherwise, if they were about the person of which they are true, they would be true after all! Hence, in using the first person (either in speech or in thought) in entertaining those beliefs, the subject must indeed be referring to, or thinking about herself, and, to that extent, she must know which person she is. What the subject would then be doing, as Christofidou claims, is to misascribe a given property to *herself*, namely, either the property *F*, or the property of being identical to *F*.

Nouce, however, that this consequence may in fact be blocked by Rovane, by

Moreover, the other proposals, according to which knowledge of which person one is is either given by the belief expressible as "I am the person who is speaking now," or by the belief expressible as "I am the thinker of this thought," face two immediate problems. First, any suggestion that construes knowledge of oneself in terms of *beliefs about oneself* presupposes a grasp of the first-person concept already, hence knowledge of which person one is. Therefore, it cannot be an explanation of what that knowledge amounts to. Second, such a suggestion would not have the resources to discriminate between the subject and its possible doppelgänger. As a consequence, the very possibility of genuine, singular I-thoughts would be impaired.[27]

Thus, I take it that Christofidou is right in thinking of knowledge of oneself as essentially unmediated by any belief about oneself and as depending instead on some kind of *acquaintance with oneself*. In fact, Evans, followed by McDowell and later by Peacocke, construed knowledge of oneself as a kind of acquaintance with one's own body, its environment, one's memories and one's own mental states. Although this is not the place to rehearse the details of their respective proposals, the important point about their views is just this: knowledge of which person one is consists in the *dispositional* ability to form thoughts about oneself on the basis of the *information* deriving from one's body, its environment, one's memory, and from one's own mental states. Hence, although the amnesiac would not actually be able to form I-thoughts on the basis of her memories, nor would an anaesthetised subject actually be able to form them on the basis of her proprioceptive feelings, or of the perception of the environment around her, they may have nevertheless retained the relevant dispositions.[28] These, in turn, once combined with one's ability to form I-thoughts on the basis of one's own occurrent mental states, would suffice for giving those subjects knowledge of which persons they are. Knowledge of which persons they are would then secure that any (comprehending) use of the first person either in speech or in thought would possess the real guarantee.

---

holding that, no matter whether one has massively false beliefs about oneself, one would still know which person one is if one had the presumably unimpaired beliefs expressible as "I am the person who is speaking now," and "I am the thinker of this thought."

[27] I would like to thank Rovane for generously drawing this latter problem to my attention.

[28] As L. O'Brien notices: "One should not conflate a non-activated disposition with the absence of that disposition at all"—"Evans on Self-Identification," *Noûs*, XXIX, 2 (1995): 232–47, here see p. 237.

V. CONCLUSIONS

Anscombe, Rovane, and Christofidou have all been after the idea that in our use of the first person, we possess some kind of distinctive and sure-fire guarantee. Yet, on reflection it has become apparent that it cannot be the case that that guarantee is either against error through misidentification, or against the possibility of a split between speaker's and semantic reference, for it is enough for a particular use of 'I' to be affected by the former error that that pronoun is used to give expression to a judgment about oneself which is based on a belief in a wrong identification component. And it is enough for a particular use of 'I' to be affected by the latter kind of mistake that one has the intention primarily to refer to a person somehow demonstratively presented to one, and wrongly taken as oneself.[29]

The real guarantee possessed by our uses of the first person is different, and more simple. It is the guarantee that we each of us know which person is the semantic referent of our tokenings of 'I' and which person our respective first-person concepts are concepts of. It is this guarantee that Rovane and Christofidou are aiming to account for, with their respective proposals about the nature of knowledge of oneself, under the misapprehension, which can be traced back to Anscombe, that it is what others have called immunity to error through misidentification. But we can set that misapprehension aside. The fundamental issue which their accounts should be seen as addressing is the epistemological nature of self-reference, whether in speech or in thought, and how it underwrites the real guarantee.

ANNALISA COLIVA

University of Bologna

[29] Notice that the difference between these two kinds of error is exhausted by the kind of (primary) referential intentions involved. Namely, the (primary) intention of saying something about oneself in the former case, and the (primary) intention of saying something about the person visually presented to one, in the latter case.

NEW BOOKS: ANTHOLOGIES

MARTIN, RAYMOND, and JOHN BARRESI, eds.: *Personal Identity*. Malden, MA: Blackwell, 2003. xiv, 389 p. Paper $29.95.

MATUŠTÍK, MARTIN BECK, and WILLIAM L. MCBRIDE, eds.: *Calvin O. Schrag and the Task of Philosophy after Postmodernity*. Northwestern: University Press, 2002. xix, 339 p. Cloth $89.95, paper $27.95.

MENDIETA, EDUARDO, ed.: *Latin American Philosophy: Currents, Issues, Debates*. Bloomington: Indiana UP, 2003. vi, 218 p. Cloth $49.95, paper $22.95.

MILLER, ALEXANDER, and CRISPIN WRIGHT, eds.: *Rule-following and Meaning*. Ithaca, NY: McGill-Queen's, 2002. x, 301 p. Cloth $75.00.

MCCORMICK, JOHN P., ed.: *Confronting Mass Democracy and Industrial Technology: Political and Social Theory from Nietzsche to Habermas*. Durham, NC: Duke, 2002. v, 368 p. Cloth $64.95, paper $21.95.

MEREDYTH, DENISE, and JEFFREY MINSON, eds.: *Citizenship and Cultural Policy*. Thousand Oaks, CA: Sage, 2001. xxx, 166 p. Cloth $81.00, paper $24.95.

MORAN, DERMOT, and TIMOTHY MOONY, eds.: *The Phenomenology Reader*. New York: Routledge, 2002. x, 614 p. Paper $27.99.

MORENO, JONATHAN D., ed.: *In the Wake of Terror: Medicine and Morality in a Time of Crisis*. Cambridge: MIT, 2003. xxv, 229 p. Cloth $24.95.

MOSER, PAUL K., ed.: *The Oxford Handbook of Epistemology*. New York: Oxford, 2002. 595 p. Cloth $74.00.

NADLER, STEVEN, ed.: *A Companion to Early Modern Philosophy*. Malden, MA: Blackwell, 2002 ix, 661 p. Cloth $124.95.

NEILL, ALEX, and AARON RIDLEY, eds.: *Arguing about Art: Contemporary Philosophical Debates*, Second Edition. New York: Routledge, 2002. xiii, 479 p. Paper $27.95.

NELSON, TODD D., ed.: *Ageism: Stereotyping and Prejudice against Older Persons*. Cambridge: MIT, 2002. xvi, 372 p. Cloth $45.00.

NICKLES, THOMAS, ed.: *Thomas Kuhn*. New York: Cambridge, 2003. xi, 298 p. Cloth $58.00, paper $21.00.

NOË, ALVA, and EVAN THOMPSON, eds.: *Vision and Mind: Selected Readings in the Philosophy of Perception*. Cambridge: MIT, 2002. x, 627 p. Paper $38.00.

PATTERSON, DENNIS, ed.: *An Anthology of Philosophy of Law and Legal Theory*. Malden, MA: Blackwell, 2003. viii, 424 p. Paper $84.95.

PAUER-STUDER, HERLINDE, ed.: *Constructions of Practical Reason: Interviews on Moral and Political Philosophy*. Stanford: University Press, 2003. xiii, 230 p. Cloth $55.00, paper $19.95.

PEDDLE, DAVID G., and NEIL G ROBERTSON, eds.: *Philosophy and Freedom: The Legacy of James Doull*. Toronto: University Press, 2003. xxix, 520 p. Cloth $115.00.

PETERS, MICHAEL A., and PAULO GHIRALDELLI, JR., eds.: *Richard Rorty: Education, Philosophy, and Politics*. Lanham, MD: Rowman and Littlefield, 2001. vi, 210 p. Cloth $67.00, paper $26.95.

PORTNER, PAUL, and BARBARA H. PARTEE, eds.: *Formal Semantics: The Essential Readings*. Malden, MA: Blackwell, 2002. x, 486 p. Paper $39.95.

PRESTON, JOHN, and MARK BISHOP, eds.: *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*. New York: Oxford, 2002. xvi, 410 p. $24.95.

PREYER, GERHARD, and GEORG PETER, eds.: *Logical Form and Language.* New York: Oxford, 2002. viii, 512 p. Paper $29.95.

RADDER, HANS, ed.: *The Philosophy of Scientific Experimentation.* Pittsburgh: University Press, 2003. xii, 311 p Paper $29.95.

RAVVEN, HEIDI M., and LENN E. GOODMAN, eds.: *Jewish Themes in Spinoza's Philosophy.* Albany: SUNY, 2002. ix, 290 p. Cloth $78.50, paper $26.95.

RICHARD, MARK, ed.: *Meaning.* Malden, MA: Blackwell, 2003. x, 341 p. Paper $29.95.

ROBINSON, THOMAS M., and LAURA WESTRA, eds.: *Thinking about the Environment: Our Debt to the Classical and Medieval Past.* Lanham, MD: Lexington, 2002. xii, 226 p. Cloth $70.00.

RORTY, AMELIE OKSENBERG, ed.: *The Many Faces of Philosophy: Reflections from Plato to Arendt.* New York: Oxford, 2003. xxix, 512 p. Cloth $40.00.

SARAT, AUSTIN, and THOMAS R. KEARNS, eds.: *Human Rights: Concepts, Contests, and Contingencies.* Ann Arbor: Michigan UP, 2002. 124 p. Cloth $39.50, paper $17.95.

SAUL, NICHOLAS, ed.: *Philosophy and German Literature, 1700–1990.* New York: Cambridge, 2002. xii, 324 p. Cloth $65.00.

SCHARFF, ROBERT C., and VAL DUSEK, eds.: *Philosophy of Technology: The Technological Condition, An Anthology.* Malden, MA: Blackwell, 2003. xi, 686 p. Paper $44.95.

SCHEMAN, NAOMI, and PEG O'CONNOR, eds.: *Feminist Interpretations of Ludwig Wittgenstein.* University Park: Penn State UP, 2002. xiii, 472 p. Cloth $85.00, paper $34.50.

SCHEUTZ, MATTHIAS, ed.: *Computationalism: New Directions.* Cambridge: MIT, 2002. xiii, 209 p. Cloth $35.00.

SCHMIDTZ, DAVID, ed.: *Robert Nozick.* New York: Cambridge, 2002. x, 230 p. Cloth $60.00, paper $20.00.

SCHWARTZ, ROBERT, ed.: *Perception.* Malden, MA: Blackwell, 2004. xii, 361 p. Paper $29.95.

SCHOLL, BRIAN J, ed.: *Objects and Attention.* Cambridge: MIT, 2002. 220 p. Cloth $40.00.

SCOTT, GARY ALAN, ed.: *Does Socrates Have a Method?: Rethinking the Elenchus in Plato's Dialogues and Beyond.* University Park: Penn State UP, 2002. xiii, 327 p. Cloth $45.00.

SHAPIRO, MICHAEL, ed.: *The Peirce Seminar Papers, Volume 5: Essays in Semiotic Analysis.* New York: Berghahn, 2002 v, 226 p. Cloth $69.95.

SHIELDS, CHRISTOPHER, ed.: *The Blackwell Guide to Ancient Philosophy.* Malden, MA: Blackwell, 2003. xi, 333 p. Paper $34.95.

SIMONS, JON, ed.: *From Kant to Levi-Strauss: The Background to Contemporary Critical Theory* Edinburgh: University Press, 2002. x, 256 p. Paper $24.00.

SINNOTT-ARMSTRONG, WALTER, and ROBERT AUDI, eds.: *Rationality, Rules, and Ideals: Critical Essays on Bernard Gert's Moral Theory.* Lanham, MD: Rowman and Littlefield, 2002. vii, 326 p. Cloth $75.00, paper $24.95.

SMITH, BARRY, ed.: *Liberal Education in a Knowledge Society.* Chicago: Open Court, 2002. viii, 291 p. Cloth $62.95, paper $28.95.

SOBLE, ALAN, ed.: *The Philosophy of Sex: Contemporary Readings,* Fourth Edition.

Lanham, MA: Rowman and Littlefield, 2002. xlii, 513 p. Cloth $90.00, paper $29.95.

SOLOMON, ROBERT C., ed.: *What Is an Emotion? Classic and Contemporary Readings*, Second Edition. New York: Oxford, 2003. x, 305 p. Paper $29.95.

SOLOMON, ROBERT C., and DAVID SHERMAN, eds.: *The Blackwell Guide to Continental Philosophy*. Malden, MA: Blackwell, 2003. viii, 345 p. Paper $34.95.

STEELE, DAVID RAMSAY, ed.: *Genius: In Their Own Words: The Intellectual Journeys of Seven Great Twentieth-Century Thinkers*. Chicago: Open Court, 2002. xxii, 366 p. Paper $24.95

SUPERSON, ANITA M., and ANN E. CUDD, eds.: *Theorizing Backlash: Philosophical Reflections on the Resistance to Feminism*. Lanham, MD: Rowman and Littlefield, 2002. xxiii, 269 p. Cloth $75.00, paper $24.95.

TALIAFERRO, CHARLES, and PAUL J. GRIFFITHS, eds.: *Philosophy of Religion: An Anthology*. Malden, MA: Blackwell, 2003. xi, 622 p. Paper $39.95.

THOMSON, JUDITH JARVIS, et al.: *Goodness and Advice*. AMY GUTMANN, ed. Princeton: University Press, 2001. xvi, 188 p. Paper $14.95.

TRAPPI, ROBERT, PAOLO PETTA, and SABINE PAYR, eds.: *Emotions in Humans and Artifacts*. Cambridge: MIT, 2003. vii, 390 p. Cloth $50.00.

TROYER, JOHN, ed.: *The Classical Utilitarians: Bentham and Mill*. Indianapolis: Hackett, 2003. xxix, 281 p. Cloth $27.95, paper $9.95.

TUANA, NANCY, WILLIAM COWLING, MAURICE HAMINGTON, GREG JOHNSON, and TERRANCE MACMULLAN, eds.: *Revealing Male Bodies*. Bloomington: Indiana UP, 2002. xi, 310 p. Cloth $62.50, paper $20.95.

TURNER, STEPHEN P., and PAUL A. ROTH, eds.: *The Blackwell Guide to the Philosophy of the Social Sciences*. Malden, MA: Blackwell, 2003. viii, 382 p. Cloth $69.95, paper $34.95.

UPHAM, S. PHINEAS, ed.: *Philosophers in Conversation: Interviews from The Harvard Review of Philosophy*. New York: Routledge, 2002. xiii, 189 p. Cloth $85.00, paper $21.95.

VAN GELDEREN, MARTIN, and QUENTIN SKINNER, eds.: *Republicanism: A Shared European Heritage, Volume 1: Republicanism and Constitutionalism in Early Modern Europe*. New York: Cambridge, 2002. xi, 420 p. Cloth $60.00; *Republicanism: A Shared European Heritage, Volume 2: The Values of Republicanism in Early Modern Europe*. New York: Cambridge, 2002. xi, 402 p. Cloth $60.00.

WILKERSON, WILLIAM S., and JEFFREY PARIS, eds.: *New Critical Theory: Essays on Liberation*. Lanham, MD: Rowman and Littlefield, 2001. xvi, 272 p. Cloth $75.00, paper $26.95.

WILLIAMS, THOMAS, ed.: *The Cambridge Companion to Duns Scotus*. New York: Cambridge, 2003. xvi, 408 p. Cloth $65.00, paper $23.00.

WOLFREYS, JULIAN, ed.: *The Continuum Encyclopedia of Modern Criticism and Theory*. New York: Continuum, 2002. xiii, 882 p. Cloth $175.00.

YANCY, GEORGE, ed.: *The Philosophical I: Personal Reflections on Life in Philosophy*. Lanham, MD: Rowman and Littlefield, 2002. xxix, 295 p. Paper $27.95.

YU, JIYUAN, and JORGE J.E. GRACIA, eds.: *Rationality and Happiness: From the Ancients to the Early Medievals*. Rochester, NY: University Press, 2003. xi, 252 p. Cloth $75.00.

NEW BOOKS

ANDRE, JUDITH: *Bioethics as Practice.* Chapel Hill: North Carolina UP, 2002. xii, 253 p. Cloth $29.95.

ANDREWS, PETER B.: *An Introduction to Mathematical Logic and Type Theory: To Truth through Proof,* Second Edition. Norwell, MA: Kluwer, 2002. xviii, 390 p. Cloth $81.00, paper $49.00.

ANGLE, STEPHEN C.: *Human Rights and Chinese Thought: A Cross-Cultural Inquiry.* New York: Cambridge, 2002. xviii, 285 p. Cloth $65.00, paper $23.00.

ANIDJAR, GIL: *The Jew, the Arab: A History of the Enemy.* Stanford: University Press, 2003. xxv, 265 p. Cloth $55.00, paper $21.95.

ARGYROU, VASSOS: *Anthropology and the Will to Meaning: A Postcolonial Critique.* Sterling, VA: Pluto, 2002. v, 129 p. Cloth $69.95, paper $24.95.

ARKES, HADLEY: *Natural Rights and the Right to Choose.* New York: Cambridge, 2002. xiii, 302 p. Cloth $28.00

ARPALY, NOMY: *Unprincipled Virtue: An Inquiry into Moral Agency.* New York: Oxford, 2003. viii, 203 p. Cloth $45.00.

ARSIC, BRANKA: *The Passive Eye: Gaze and Subjectivity in Berkeley (via Beckett).* Stanford: University Press, 2003. xiii, 210 p. Cloth $45.00, paper $19.95.

ATRAN, SCOTT: *In Gods We Trust: The Evolutionary Landscape of Religion.* New York: Oxford, 2002. xvi, 348 p. Cloth $35.00.

BAGGINI, JULIAN: *Making Sense: Philosophy behind the Headlines.* New York: Oxford, 2002. 296 p. Cloth $26.00; *Philosophy: Key Texts.* New York: Palgrave Macmillan, 2002. xii, 140 p. Cloth $65.00, paper $19.95; *Philosophy: Key Themes.* New York: Palgrave Macmillian, 2002. xi, 140 p. Cloth $65.00, paper $19.95.

BAGGINI, JULIAN, and PETER S. FOSL: *The Philosopher's Toolkit: A Compendium of Philosophical Concepts and Methods.* Malden, MA: Blackwell, 2003. x, 221 p. Paper $19.95.

BALFOUR, IAN: *The Rhetoric of Romantic Prophecy.* Stanford: University Press, 2002. x, 346 p. Cloth $65.00, paper $24.95.

BANDMAN, BERTRAM: *Children's Right to Freedom, Care, and Enlightenment.* New York: Garland, 1999. xxxi, 366 p. Cloth $65.00.

BARNARD, F.M.: *Herder on Nationality, Humanity, and History.* Ithaca: McGill-Queen's, 2003. xii, 185 p. Cloth $70.00.

BEHNEGAR, NASSER: *Leo Strauss, Max Weber, and The Scientific Study of Politics.* Chicago: University Press, 2003. xiii, 221 p. Cloth $30.00.

BENHABIB, SEYLA: *The Claims of Culture: Equality and Diversity in the Global Era.* Princeton: University Press, 2002. xvii, 245 p. Cloth $49.95, paper $16.95.

BENNETT, M. R., and P. M. S. HACKER: *Philosophical Foundations of Neuroscience.* Malden, MA: Blackwell, 2003. xvi, 461 p. Cloth $74.95, paper $39.95.

BENSON, BRUCE ELLIS: *Improvisation of Musical Dialogue.* New York. Cambridge, 2003. xiv, 200 p. Cloth $53.00, paper $19.00.

BERGER, ALAN: *Terms and Truth: Reference Direct and Anaphoric.* Cambridge: MIT, 2002. xvii, 234 p. Cloth $35.00.

BERMÚDEZ, JOSÉ LUIS: *Thinking without Words.* New York: Oxford, 2003. xiii, 225 p. Cloth $35 00.

BHASKAR, ROY: *Meta-Reality: The Philosophy of Meta-Reality, Volume 1: Creativity, Love and Freedom.* Thousand Oaks, CA: Sage, 2002. liv, 364 p. Paper $24.95.

BOBONICH, CHRISTOPHER: *Plato's Utopia Recast: His Later Ethics and Politics.* New York: Oxford, 2002. xi, 643 p. Cloth $49.95.

BONEVAC, DANIEL: *Deduction: Introductory Symbolic Logic,* Second Edition. Malden, MA: Blackwell, 2003. x, 514 p. Paper $49.95.

BONJOUR, LAURENCE, and ERNEST SOSA: *Epistemic Justification: Internalism versus Externalism, Foundations versus Virtues.* Malden, MA: Blackwell, 2003. vii, 240 p. Paper $26.95.

BOONIN, DAVID: *A Defense of Abortion.* New York: Cambridge, 2003. xvi, 350 p. Cloth $65.00, paper $23.00.

BORRADORI, GIOVANNA: *Philosophy in a Time of Terror: Dialogues with Jürgen Habermas and Jacques Derrida.* Chicago: University Press, 2003. xvi, 208 p. Cloth $25.00.

BOSTROM, NICK: *Anthropic Bias: Observation Selection Effects in Science and Philosophy.* New York: Routledge, 2002. xiii, 224 p. Cloth $70.00.

BRADY, EMILY: *Aesthetics of the Natural Environment.* Tuscaloosa: Alabama UP, 2003. viii, 287 p. Cloth $67.50, paper $26.95.

BRAMS, STEVEN J.: *Biblical Games: Game Theory and the Hebrew Bible.* Cambridge: MIT, 2003. xv, 220 p. Paper $17.95.

BRANDIST, CRAIG: *The Bakhtin Circle: Philosophy, Culture and Politics.* Sterling, VA: Pluto, 2002. x, 221 p. Cloth $69.95, paper $24.95.

BRANDOM, ROBERT B.: *Tales of the Mighty Dead.* Cambridge: Harvard, 2002. x, 430 p. Cloth $39.95.

BRINGSJORD, SELMER, and MICHAEL ZENZEN: *Superminds: People Harness Hypercomputation, and More.* Norwell, MA: Kluwer, 2003. xxx, 341 p. Cloth $125.00.

BRUNO, GIULIANA: *Atlas of Emotion: Journeys in Art, Architecture, Film.* New York: Verso, 2002. xi, 484 p. Cloth $39.00.

CAHN, STEVEN M.: *Puzzles and Perplexities: Collected Essays.* Lanham, MD: Rowman and Littlefield, 2002. ix, 137 p. Cloth $70.00, paper $22.95.

CAMPBELL, JOHN: *Reference and Consciousness.* New York: Oxford, 2002. vii, 267 p. Paper $19.95.

CARD, CLAUDIA: *The Atrocity Paradigm: A Theory of Evil.* New York: Oxford, 2002. x, 284 p. Cloth $39 95.

CASULLO, ALBERT: *A Priori Justification.* New York: Oxford, 2003. xiii, 249 p. Cloth $55.00.

CHADWICK, DANIEL R.: *Tables for Students of Philosophy: A Supplement to Philosophical Discussion.* Lanham, MD: UP of America, 2002. v, 187 p. Paper $70.00.

CHATTERJEA, TARA: *Knowledge and Freedom in Indian Philosophy.* Lanham, MD: Lexington, 2002. xvi, 159 p. Cloth $60.00.

CHEVALIER, JACQUES M.. *Half Brain Fables and Figs in Paradise: The 3-D Mind, Volume One.* Ithaca, NY: McGill-Queen's UP, 2002. 197 p. Cloth $34.95; *The Corpus and the Cortex: The 3-D Mind, Volume Two.* Ithaca, NY: McGill-Queen's UP, 2002. 288 p. Cloth 34.95, *Scorpions and the Anatomy of Time: The 3-D Mind, Volume Three.* Ithaca, NY: McGill-Queen's, 2002. 224 p. Cloth $34.95.

THE JOURNAL OF PHILOSOPHY

SUBSCRIPTIONS

2003

Students $20.00
Individuals $35.00
Institutions $75.00
Postage outside the U.S. $15.00

SPECIAL OFFER

one complimentary copy of any of the following special issues with each new individual subscription

2000	*Causation*	David Lewis, Ned Hall, John Collins, L.A. Paul, Jonathan Schaffer
1998	*Having the World in View: Sellars, Kant, and Intentionality*	John McDowell
1995	*Reconciliation through the Public Use of Reason*	Jürgen Habermas
	Reply to Habermas	John Rawls
1994	*Sense, Nonsense, and the Senses: An Inquiry into the Powers of the Human Mind*	Hilary Putnam
1990	*The Structure and Content of Truth*	Donald Davidson
1985	*Well-Being, Agency, and Freedom*	Amartya Sen
1980	*Kantian Constructivism in Moral Theory*	John Rawls
1977	*Teleology Revisited*	Ernest Nagel
1968	*Ontological Relativity*	W.V. Quine

All issues above are also available for $8.00 plus $3.00 for shipping/handling ($5.00 for international and Canadian orders). Postage for each added issue, $1.00. Send check or money order to:

Journal of Philosophy
709 Philosophy Hall
1150 Amsterdam Avenue, MC 4972
Columbia University
New York, NY 10027
(212) 666-4419
www.journalofphilosophy.org

OUTSTANDING SCHOLARSHIP FROM CAMBRIDGE

Aquinas: *Political Writings*
Edited by R.W. Dyson
Cambridge Texts in the History of Political Thought
0-521-37595-9 Paperback $22.00

Ethics and Foreign Intervention
Edited by Deen K. Chatterjee and Don E. Scheid
Cambridge Studies in Philosophy and Public Policy
0-521-81074-4 Hardback $58.00
0-521-00904-9 Paperback $21.00

Towards a Philosophy of Real Mathematics
David Corfield
0-521-81722-6 Hardback $65.00

Meaning in Spinoza's Method
Aaron V. Garrett
0-521-82611-X Hardback $60.00

The Cambridge Companion to Pascal
Edited by Nicholas Hammond
Cambridge Companions to Philosophy
0-521-80924-X Hardback $60.00
0-521-00611-2 Paperback $23.00

Donald Davidson
Edited by Kirk Ludwig
Contemporary Philosophy in Focus
0-521-79043-3 Hardback $60.00
0-521-79382-3 Paperback $20.00

Beyond Comparison: Sex and Discrimination
Timothy Macklem
Cambridge Studies in Philosophy and Law
0-521-82682-9 Hardback $65.00
0-521-53415-1 Paperback $26.00

Alasdair MacIntyre
Edited by Mark C. Murphy
Contemporary Philosophy in Focus
0-521-79042-5 Hardback $65.00
0-521-79381-5 Paperback $20.00

Logic, Thought and Language
Edited by Anthony O'Hear
Royal Institute of Philosophy Supplements 51
0-521-52968-2 Paperback $20.00

A Question of Trust
The BBC Reith Lectures 2002
Onora O'Neill
0-521-82304-8 Hardback $40.00
0-521-52996-4 Paperback $15.00

After Socialism
Edited by Ellen Frankel Paul, Fred D. Miller, Jr., and Jeffrey Paul
Social Philosophy and Policy 20
0-521-53498-4 Paperback $25.00

The Difficulty of Tolerance
Essays in Political Philosophy
T.M. Scanlon
0-521-82661-6 Hardback $55.00*
0-521-53398-8 Paperback $20.00*

The Cambridge Companion to Greek and Roman Philosophy
Edited by David Sedley
Cambridge Companions to Philosophy
0-521-77285-0 Hardback $65.00*
0-521-77503-5 Paperback $24.00*

Visions of Politics
Volumes 1-3
Quentin Skinner
0-521-89075-6 Paperback Set $65.00

The Uses of Argument
Updated Edition
Stephen E. Toulmin
0-521-82748-5 Hardback $55.00
0-521-53483-6 Paperback $20.00

Veil Politics in Liberal Democratic States
Ayume H. Wingo
Cambridge Studies in Philosophy and Public Policy
0-521-81438-3 Hardback $55.00
0-521-89128-0 Paperback $22.00

Thomas Reid and the Story of Epistemology
Nicholas Wolterstorff
Modern European Philosophy
0-521-53930-7 Paperback $24.00*

Available in bookstores or from

*Prices subject to change

CAMBRIDGE UNIVERSITY PRESS us.cambridge.org/philosophy

New from
Oxford







Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556. Visit our website at www.oup-usa.org

OXFORD

Reproductive Health and Human Rights
Integrating Medicine, Ethics, and Law

REBECCA J. COOK, BERNARD M. DICKENS, and MAHMOUD F. FATHALLA

The authors analyze fifteen case-studies, representing a wide array of recurrent problems, focusing particularly on resource-poor settings. Approaches to resolution are considered at clinical and health system levels. They also consider the kinds of social change that would relieve the underlying conditions of reproductive health dilemmas.

2003 paper $29.95 cloth $95.00

New in Paperback!

Science, Truth, and Democracy

PHILIP KITCHER

"Kitcher navigates very skillfully between the extremes of positivistic science-worship and Foucauldian distrust of 'regimes of truth.' His novel and plausible answer to the question 'Why seek scientific truth?' will help bring the increasingly tedious Science Wars to a close.—Richard Rorty, Stanford University

paper September 2003 $17.95

The Seventh Sense
Frances Hutcheson and Eighteenth-Century British Aesthetics

PETER KIVY

Now reissued with substantial new material, *The Seventh Sense* is the definitive study of the aesthetic theory of the great eighteenth-century philosopher Frances Hutcheson, and its huge influence on British aesthetics.

2003 paper $24.95 cloth $74.00

New in Paperback!

Facing Facts

STEPHEN NEALE

"Neale's book is written with such thoroughness, clarity, and rigor...No one with an interest in intensionality or the semantics of descriptions will want to miss this book."—John MacFarlane, *Notre Dame Philosophical Reviews*

paper September 2003 $19.95

New in Paperback!

Nature and Understanding
The Metaphysics and Methods of Science

NICHOLAS RESCHER

This book seeks to describe in a clear, accessible manner the metaphysical situation that characterizes the process of inquiry in natural science, aiming to shed light on reality by examining the modus operandi of natural science itself and focusing as much on its findings as on its conceptual and methodological presuppositions.

paper 2003 $22.00

New in Paperback!

Four-Dimensionalism
An Ontology of Persistence and Time

THEODORE SIDER

Lucid, absorbing, and original, this book defends the theory that the material world is composed of temporal as well as spatial parts. Along the way, it addresses many topics on the metaphysics of time and identity.

paper 2003 $19.95

THE BRITISH JOURNAL OF

Aesthetics

print • 4 issues a year • online

www.bjaesthetics.oupjournals.org

The British Journal of Aesthetics is published to promote study, research, and discussion of the fine arts and related types of experience from a philosophical, psychological, sociological, scientific, historical, critical, and educational standpoint.

Founded in 1960, *The British Journal of Aesthetics* is highly regarded as an international forum for debate in aesthetics and the philosophy of art. It publishes lively and thoughtful articles on a broad range of topics from the nature of aesthetic judgement and the principles of art criticism to foundational issues concerning the visual arts, literature, music, dance, film, and architecture.

Visit the philosophy website to:

- ❒ Receive tables-of-contents by e-mail
- ❒ Browse tables-of-contents FREE of charge
- ❒ Place a subscription
- ❒ View free sample copies online

www.oupjournals.org/philosophy

OXFORD
UNIVERSITY PRESS

International Journal of Philosophical Studies

EDITOR
Dermot Moran, *University College Dublin, Ireland*

REVIEW EDITOR
Maria Baghramian, *University College Dublin, Ireland*

The *International Journal of Philosophical Studies* publishes academic articles of the highest quality in all areas of philosophy In particular the journal seeks to

promote mutual comprehension and dialogue between the 'analytic' and 'continental' styles of philosophy
combine discussion of contemporary problems with articles in the history of philosophy which shed light on aspects of the current debate
provide quality articles on a broader range of issues than is currently available in philosophical journals

The journal contains an extensive book review section, including reviews of non-English language books It also provides Critical Notices which review major books or themes in depth

This journal is also available online. Please connect to www.tandf.co.uk/online.html for further information.

To request a sample copy please visit www.tandf.co.uk/journals

SUBSCRIPTION RATES
2003 – Volume 11 (4 issues)
Print ISSN 0967-2559 Online ISSN 1466-4542
Institutional rate US$381, £230
(includes free online access)
Personal rate US$107, £65 (print only)

Routledge
Taylor & Francis Group

ORDER FORM riph

PLEASE COMPLETE IN BLOCK CAPITALS AND RETURN TO THE ADDRESS BELOW

Please invoice me at the ❑ institutional rate ❑ personal rate

Name _____

Address _____

_____

_____

Email _____

Graduate Faculty Philosophy Journal

The **Graduate Faculty Philosophy Journal**, published in association with the Department of Philosophy, New School for Social Research, is a forum for the communication of ideas in the history of philosophy and its interpretation

VOL. 24 NO. 1

BURT HOPKINS — Crisis, History, Husserl and Jacob Klein's Contribution

JULIE KLEIN — Spinoza's Debt to Gersonides

ALEXANDRE KOYRÉ — Paracelsus (1493-1541)

HANS KRÄMER — The Concept of Hermeneutical Experience: Alternatives to Gadamer

WALLACE MATSON — Zombies Begone! Against Chalmers' Mind/Brain Dualism

MONIQUE DAVID-MENARD — Heterogeneous Disciplines: A Non-Ontological Strategy

RUPERT READ — Kripke's Hume

PIRMIN STEKELER-WEITHOFER — Noumenal Will in Kant's Theory of Action

MORAN SVORAI — On the Philosophical Significance of Transference

REVIEW ESSAYS

BERNARD FLYNN — *Against Adaptation: Lacan's Subversion of the Subject* by Philippe van Haute

NECTARIOS LIMNATIS — *German Philosophy 1760-1860: The Legacy of Idealism* by Terry Pinkard

All communications should be addressed to the Editor, **Graduate Faculty Philosophy Journal**, Department of Philosophy, New School for Social Research, 65 Fifth Avenue, New York, NY 10003. The Journal is biannual. Domestic rates: Individuals: $20.00/year; Students: $12.00/year; Institutions: $45.00/year.



APA members save 20%

Purpose and Scope

Founded in 1890, ***Ethics*** is an international journal of moral, political, and legal philosophy. ***Ethics*** presents scholarly work from a variety of intellectual perspectives, including philosophy, social and political theory, law, and economics. The journal presents theory and its applications on contemporary moral issues. In addition, ***Ethics*** offers about 200 review essays, survey articles, book reviews, and book notes per volume.

Special Features

Ethics also holds symposia and publishes special issues that examine particular topics in depth through the contribution of distinguished scholars.

Ethics Electronic Edition

- Links to author references, tables, and figures
- Articles available for PDF printing capability
- Personalized passwords allow flexible access from any location

FREE Table of Contents Alerting

Sign up for a free table of contents alerting, please visit
www.journals.uchicago.edu/ET/journal/index.html

The University of Chicago Press



THE JOURNAL OF PHILOSOPHY

Celebrates 100 Years

Special Limited Time Offer

Full Volumes 1-99

1904-2002

$50.00 per Volume

All volumes are available for $50.00 plus $5.00 shipping and handling.

Send check or money order to:

The Journal of Philosophy

1150 Amsterdam Avenue, MC 4972

Columbia University

New York, NY 10027

(212) 666-4419

www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY



VOLUME C, NUMBER 9
SEPTEMBER 2003

page

437 *The Ordinary Concept of Race* — Michael O. Hardimon
456 *On the New Biology of Race* — Joshua M. Glasgow

COMMENTS AND CRITICISM

475 *Thomson and the Semantic Argument against Consequentialism* — David Phillips

BOOK REVIEWS

487 Henry E. Allison: *Kant's Theory of Taste: A Reading of the Critique of Aesthetic Judgment* — Béatrice Longuenesse

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX: (212) 932-8721.

You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2003

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset and printed by Capital City Press, Montpelier, VT.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

POSTMASTER: Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2003 by the Journal of Philosophy, Inc.

ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY
VOLUME C, NO. 9, SEPTEMBER 2003

THE ORDINARY CONCEPT OF RACE*

The ordinary concept of race is important and poorly understood. The present article seeks to address this problem by providing a general answer to the question: What is the concept of race?[1]

I

The ordinary concept of race corresponds (roughly) to the meaning of the ordinary word 'race'. Deeply entrenched in our social practices, it is the concept competent users of the word 'race' use the word to express. The concept plays an especially central role in our *racialized* social practices: practices which cannot be identified or explained without reference to the concept of race such as segregation, racial profiling, and affirmative action.

The ordinary concept of race is ordinary. It is used to categorize common perceived objects—human beings and their properties—in the course of everyday life and can be applied through the use of ordinary perceptual powers. The ordinary term 'race' bears the imprint of eighteenth- and nineteenth-century attempts to develop a scientific concept of race (race as an infraspecific division of *Homo sapiens*), which was an outgrowth of the Enlightenment's larger taxonomic project. But the ordinary word never left ordinary discourse.

* I would like to express thanks to Mary Devereaux for helpful comments.

[1] The present essay is concerned with the "meaning" (or content) of the ordinary concept of race. It does not, however, address some issues that are commonly associated with talk about the "significance" of race. Thus, for example, it does not discuss the *experience* of being a member of a particular racial group. It seeks to get clear about *what it is* we are asking about *when* we ask about the "significance of being _____" (black, Latino, Asian, and so forth). I have taken the liberty of helping myself to the familiar names of "racial" groups. I recognize that these names are contested and that the same is true of race names generally.

0022-362X/03/0009/437–55 © 2003 The Journal of Philosophy, Inc.

The ordinary concept has a long established life *outside* of science and deserves to be investigated on its own.[2]

What precisely the relation between the ordinary concept of race and the scientific concept of race comes to is a matter of some controversy. A closely related dispute concerns the scientific standing (the biological respectability) of the ordinary concept of race. Both controversies are complicated by an underlying dispute concerning the *existence* of a scientific concept of race. Is there such a concept? Does biology need such a concept?[3] If so, how is it to be understood?[4] Questions such as these have dominated the discussion of the concept of race.[5] I mention them to set them to the side. Before we can sensibly address these issues, we need to get clearer about *what* the concept is.

The ordinary concept of race is *our* concept. It is part of *our* discourse, *our* practices, *our* conceptual repertoire.[6] Refusing to use the word 'race' does not acquit one of possession of the concept. It is, on the contrary, a mark of sophisticated possession. Whatever our attitude toward the concept may be, we (competent users of the English word 'race') relate to the concept as *insiders*. The 'our' is first-person *plural*. The ordinary concept of race is a concept we *share*. It is not to be confused with the sundry private "concepts" or images of race that particular individuals may have.

Why is a *philosophical* investigation of the ordinary concept of race called for? After all, the concept does not belong to the collection of concepts with which philosophy has traditionally been concerned.[7] Why not leave race to the experts—biologists, anthropologists, or social theorists—researchers whose discipline and training provides them with special expertise? To begin with, none of these disciplines are concerned with the ordinary concept of race *as such*. The examination of ordinary concepts as ordinary concepts is, however, one of

[2] Cf. Naomi Zack, *Philosophy of Science and Race* (New York: Routledge, 2002), p. 3

[3] For defense of the negative answer to these questions, see Zack.

[4] Ernst Mayr, *Animal Species and Evolution* (Cambridge: Harvard, 1963), pp. 644–50, Philip Kitcher, "Race, Ethnicity, Biology, Culture," in Leonard Harris, ed., *Racism* (New York: Humanities, 1999), pp. 87–117.

[5] See, for example, Ashley Montagu, "The Concept of Race," in Montagu, ed., *The Concept of Race* (New York: Free Press, 1964), pp. 12–28 (hereafter "Concept"); K. Anthony Appiah, "Race, Culture, Identity," in Appiah and Amy Gutmann, *Color Consciousness* (Princeton. University Press, 1996), pp. 30–105; also see Zack.

[6] The 'our' in "our concept" is not the 'our' of celebration but a simple acknowledgment of fact.

[7] Or does it? Immanuel Kant is widely regarded as the author of the first attempt to give a scientific definition of race based on a sharp distinction between race and species—see, for example, "Von der verschiedenen Rassen der Menschen," "Of the Different Human Races" (1777) in Robert Bernasconi and Tommy L. Lott, eds., *The Idea of Race* (Indianapolis: Hackett, 2000), pp. 8–26.

the traditional tasks of philosophy. Researchers in other disciplines quite properly approach race as a concept within their own fields. Biologists take the concept to be biological.[8] Social theorists take it to be social.[9] Philosophy, on the other hand, provides the resources to explore the concept without prejudging its disciplinary status.

But the primary motivation for turning to philosophy is that we are fundamentally confused about the concept of race. This confusion is illustrated by the ongoing debate over the elimination of the concept of race. Ought the concept of race be banished from our discourse and practices, as eliminativists suggest? Or, ought it be retained, at least for the present, as antieliminativists suggest? At the root of this disagreement is a dispute about how the concept of race is to be understood—a point that is often overlooked because each side assumes the other is operating with the former's conception. But *this* dispute is quintessentially philosophical. Philosophical problems call for philosophical solutions.

Before we can turn to the philosophical account of the ordinary concept of race, it is necessary to address a few preliminary matters.

(1) We need to make a philosophical distinction. The ordinary *concept* of race and the ordinary *conception* of race are two different things. The distinction between concept and conception was introduced by H.L.A. Hart[10] and gained currency through the work of John Rawls.[11] It is also found in the work of Tyler Burge.[12] In *A Theory of Justice*, Rawls employs the distinction to distinguish his preferred *conception* of justice (that is, justice as fairness) from the *concept* of justice it articulates. A full treatment of the distinction is out of the question here, but some discussion is indispensable inasmuch as the distinction plays a critical role in framing the discussion.

Concept and conception stand in a one to many relationship. It is part of the idea of a *concept* that one and the same concept can be articulated in a number of different and competing ways. It is part of the idea of a *conception* that a conception represents but one of a number of possible different and competing ways in which a given concept can be articulated. Concept and conception are correlative terms. To characterize an idea as a *concept* of X is to characterize it as the object of a possible correlative *conception* of X and conversely.

[8] Richard A. Goldsby, *Race and Races*, Second Edition (New York: Macmillan, 1977).

[9] Michael Omi and Howard Winant, *Racial Formation in the United States*, Second Edition (New York: Routledge, 1994)

[10] *The Concept of Law* (New York: Oxford, 1961), pp. 115–57.

[11] *A Theory of Justice* (Cambridge. Harvard, 1971), p. 5.

[12] "Concepts, Definitions, and Meaning," *Metaphilosophy*, IV, 4 (October 1993) · 309–23, see p. 316.

The terms 'concept' and 'conception' operate in tandem. The concept of *X* specifies what *X* is. A conception of *X* indicates how the concept of *X* is to be understood.

The concept/conception distinction makes it possible to see that much of what is commonly presented in the literature as a discussion of the ordinary *concept* of race is perhaps better understood as a discussion of the ordinary *conception* of race. The ordinary *conception* of race is what corresponds to common-sense (ordinary, everyday) ways of thinking about race (to the "layman's conception of race")[13] on which writers focus. Our focus, however, is on the ordinary *concept*.

(2) Any sensible discussion of the concept of race must start out from an awareness of what Ashley Montagu has called the "long and tortured history" of the term 'race' (*ibid.*, p. 14). It is well known that the concept has functioned historically to support racist conduct, attitudes, institutions, and practices, and that it has figured centrally in slavery, colonialism, and genocide. An acceptable account of the ordinary concept of race must also recognize that the word 'race' is emotionally loaded in a way few terms are. It is as Montagu says a "trigger word" (*ibid.*, p. 24). Henry Louis Gates, Jr. wrote poignantly at the end of the last century, "scores of people are killed every day in the name of differences ascribed only to race."[14] K. Anthony Appiah properly calls us to be ever mindful of "the evil done by the concept."[15] Clearly, the ordinary concept of race is no *ordinary* ordinary concept. This fact too must be reflected in the philosophical account.

(3) The present account does not propose to rehabilitate the word 'race'.[16] Nor does it propose to introduce a new sense of the word.[17] It works with what is already there. It takes the form of reflection on an *already given* concept. Still less is the account meant to provide warrant for the uncritical first-order use of the term. It is meant rather to facilitate *critical* reflection and to shed light on problems associated with the term. It does not engage the question whether the ordinary concept of race *ought* to be used except briefly at the end. That question can only be sensibly addressed *once* the concept has been clarified.

(4) Our project carries with it a certain ineliminable element of rationalization.[18] We want to *understand* the ordinary concept of race.

[13] Montagu, "Concept," p. 24.

[14] "Writing 'Race' and the Difference It Makes," in Gates, ed., *"Race," Writing, and Difference* (Chicago: University Press, 1986), pp. 1–20, here p. 6.

[15] Appiah, "The Uncompleted Argument," in Gates, ed., pp. 21–37, here p. 35.

[16] Cf. Montagu, "Concept," p. 15.

[17] Cf. Montagu, "Concept," pp. 19, 26; "The Concept of Race in the Human Species in the Light of Genetics," in Montagu, ed., pp. 1–11, see p. 9 (hereafter "Species").

[18] 'Rationalization' is a janus-faced word. It can be used both pejoratively and honorifically. A rationalization in the pejorative sense fails to make rational. This use is associated

But to attempt to understand is to attempt to make intelligible. Understanding *is* making intelligible. The standard approach to the concept seeks to find error, confusion, and contradiction. This essay seeks to make the best sense possible of the ordinary concept of race. It is guided by the maxim of avoiding the attribution of conceptual confusion or empirical error to the concept, wherever possible.

(5) The present account proceeds from the familiar methodological assumption that appeals to intuition are a legitimate and inescapable component of the articulation of ordinary concepts. Were the concept at issue anything other than the concept of race, the point would hardly bear mentioning. But the concept at issue *is* the concept of race. The very idea of appealing to intuitions about *it* may seem unseemly or suspect. To be sure, our intuitions about race *are* suspect. History shows how thought about race may be tainted by racist ideas in ways that are hidden to the thinker. We cannot accord the intuitions we have about the concept of race the degree of trust we give to ordinary ordinary concepts. But, again, the ordinary concept of race is *our* concept. As competent users of the term, we are *entitled* to draw on our intuitions about its application. And in any case, we could not simply ignore our intuitions even if we wanted to. They constitute our data. Even to begin the sort of inquiry undertaken here, we must suppose that at least *some* of our intuitions about the concept of race are correct. The idea that *all* our intuitions about the concept are incorrect runs into well-known Davidsonian difficulties. Surely we need to know *something* about the concept in order to hold mistaken views about *it*.[19] Still, we need to tread lightly. The possibility of ideological distortion is ever present. One small precautionary step we can take is to rely only on intuitions that survive critical reflection and remain stable in face of the evidence. These intuitions, too, may be brought into question, but they provide a place to start.

II

The aim of this section is to present the philosophical account of the concept of race. This account takes the form of an articulation of the concept's *logical core* and the *racialist development* of that core. I will discuss these two topics in turn.

II.A. The logical core of the ordinary concept of race is the concept's intelligible nucleus and is characterized by three basic theses. All

---

with the name of Freud. A rationalization in the honorific sense succeeds in making rational. In the present context, the word is to be taken in its nonpejorative sense.

[19] See Donald Davidson, "On the Very Idea of a Conceptual Scheme," in *Inquiries in Truth and Interpretation* (New York: Oxford, 1984), pp. 183–98.

three are needed to arrive at a proper understanding of the concept. They must be taken in conjunction.

The first thesis holds that

(1) The concept of race is the concept of a group of human beings distinguished from other human beings by visible physical features of the relevant kind.

This thesis captures the basic intuition that race is essentially *manifest*. The ordinary concept of race represents different racial groups as being visibly different. 'Visible' here means: visible to the naked eye, salient, and easily recognized.[20] The features in question are *bodily* features; the differences, *bodily* differences. The features are the sort of characters biologists call "phenotypic." The idea of race is the idea of a group with a distinctive phenotype.

The visibility of these features is important. Race is *essentially* manifest. The ordinary concept of race requires that the distinction between different racial groups be visibly marked in *some way or other*. The idea that there *could be* visually unmarked races—racial groups distinguished only by *invisible* features such as lactase tolerance (presence of the enzyme lactase) or the presence of the sickle-cell gene—is ruled out from the start.[21] The very notion of a visually indistinguishable racial group runs counter to the idea of race.

The insertion of the qualification 'of the relevant kind' before 'visible physical features of race' is motivated by the thought that some visible physical features are not relevant. 'Relevant' here means: relevant to counting as a racial group. This point, although rarely denied, is commonly overlooked.[22] A thought experiment will show the necessity of the proviso.

[20] Visible to the naked eye need not mean visible to the *untrained* eye. The ability to identify and distinguish visible features of the relevant kind is something that has to be *learned*.

[21] Jared Diamond proposes that human races could be defined by such "hidden" properties as the absence of anti-malarial genes (which would group Swedes together with Xhosas) and the presence of the enzyme lactase (which would group northern and central Europeans together with the Fulani of West Africa in the "lactase-positive race")—"Race and Color," *Discover* (November 1994): 84–89, here p. 86.

[22] Consider, for example, Luigi Luca Cavalli-Sforza's definition of a race as "a group of individuals that we can recognize as biologically different from others"—*Genes, People, and Languages* (New York: North Point, 2000), p. 25; Mayr's characterization of 'race': "if the average difference between two groups of individuals is sufficiently great to be recognizable on sight, we refer to such groups of individuals as races"—*Evolution and the Diversity of Life* (Cambridge: Harvard, 1975), and Montagu's representation of 'race' as "a subdivision of the species the members of which resemble each other and differ from other members of the species in certain respects"—"Concept," p. 14. I do not mean to suggest that any of these authors actually think that just any set of visible physical differences between two human groups is sufficient to determine them as different races. Presumably, they all have *populations* (that is *interbreeding* human groups) in mind when referring to visible physical differences (interbreeding groups being

Imagine a single room occupied by a racially diverse group of people. There are, let us suppose, Asian men and Asian women, white men and white women, black men and black women, and so forth. We politely ask the individuals who have Adam's apples to stand on one side of the room, and those who do not to stand on the other side. There are, let us suppose, no borderline or ambiguous cases.

The outcome of this imaginary procedure is that we have two clearly distinct groups—one on each side of the room—divided on the basis of "visible physical characteristics." One group consists of *men* (black, white, Asian), the other of *women* (black, white, Asian). Now we ask: Is this division a *racial* division?[23] I assume that the answer you will give is "no." I also assume that the 'no' is clear and unambiguous.

Interestingly, this 'no' does not appear to depend on the peculiarities of the system of racial classifications one happens to employ. It is hard to see how *any* classification we could recognize as *racial* could possibly assign men and women to distinct racial groups. One could always *call* the division between men and women "racial." But then, intuitively, the word 'race' would no longer be used to express the concept it *is* used to express. The fact that we do not think of division by Adam's apples as racial shows something important about our own understanding of the concept of race: we recognize some visible physical differences as relevant to race, others as irrelevant. Contrary to what is often said, the content of the ordinary concept of race is not completely indeterminate.

As for the *sort* of visible physical differences that are relevant, there is remarkably widespread agreement. These differences are thought to include: differences of skin color (perhaps *the* preeminent example of a *relevant* visible physical difference) and differences of facial structure, hair type, nose shape, and body form. There is no agreed upon canonical list of relevant features and no single agreed upon standard for determining what features are relevant, but there is a significant degree of overlap among lists. Everyone agrees that human beings differ along the dimensions mentioned. The differences are all differences that eliminativists are at pains not to deny.[24] And everyone

---

necessarily dimorphic) But the fact that they neglect to make this point explicit illustrates the way the point that only some visible physical differences are racial can be overlooked.

[23] The same point can be conveyed without the using the word 'racial'. We ask: Is this division the sort of division that is *called* "racial"? I assume the answer remains "no."

[24] Appiah, who thinks that "nobody has a race" finds no difficulty in speaking of "[c]lasses of people who share certain easily observable physical characteristics, most notable skin color and a few visible features of the face and head." These are "characteristics we are very good at recognizing"—"Race," pp 37, 68, and 69. Diamond, who holds that "[t]he reality of human races is another common-sense 'truth' destined to follow the flat earth into oblivion" states matter of factly, "We know that different populations classified together in the human species are visibly different"—"Race and Color," p. 88 and 86. Richard Lewontin, who, if not an eliminativist, is a celebrated critic of the

agrees that these differences are racial—or, in any case, that they are the kind of differences that are *called* racial.[25]

What I have said may be misleading so let me pause to state explicitly what it is that I am *not* saying. I am not saying that *each* racial group is distinguished from *every* other group along *each* of the dimensions specified above. Nor am I claiming that particular racial groups can be reliably identified on the basis of any one of the relevant visible differences. I am not claiming that racial groups can be distinguished on the basis of visible differences *alone*. I am not asserting that the visible physical differences provide a basis on which each individual can be assigned to a determinate racial group. Nor am I assuming that each individual belongs to exactly one racial group. I am not denying that the relevant visible physical features of race grade continuously. I am not suggesting that individual members of a racial group will be identical with respect to the visible physical features of race. I am not denying that the visible physical features of the individual members of a particular racial group will differ in various respects. My point is that racial groups differ from one another along *some* of these dimensions, and that they do so in *varying* ways and to *varying* degrees. How precisely the visible physical features of race are distributed across the individual members of a racial group, the ordinary concept of race does not say.

The specific features we have classified as "relevant" are all archetypical examples of the kind of visible physical features that are called "racial." Since our aim has been to get a grip on the sort of visible physical features that distinguish *racial* groups, this result is not unexpected. Racial groups are distinguished from one another by visible physical features that are racial. This point is rescued from tautological triviality by the fact that the relevant physical features of race can be picked out "directly" through ostension. We can say that the *relevant*

---

ordinary conception of race, says that "[o]n a broad scale, there are large and obvious differences in skin color, hair form, stature, and language from one geographical region to another"—*Human Diversity* (New York: Scientific Library, 1995), p. 116. Montagu, who holds that the word 'race' "should be permanently dethroned from the vocabulary," writes "[o]f course there exist differences"—"Concept," p. 13 and p. 27; and Zack, who calls upon us to "remove biological race from secular ontology," speaks nonchalantly of "evident human differences in skin color and morphology," p. 7 and p. 25.

[25] Holding that there are no races, some eliminativists want to deny the existence of differences that are *genuinely* racial (no race, no *racial* difference). But it seems to me that one can *call* certain properties "racial" without supposing that race is real. Having one horn is a property of unicorns despite their nonexistence. But the step *from* accepting that certain groups of individuals have characters that are properly called "racial" *to* the idea that groups whose members exhibit such features are racial, *to* the idea that *races* exist is short. Its smallness may be one source of some eliminativists' reluctance to allow that human beings have features that are properly called racial.

features are features such as *these* (pointing to skin color, eye shape, hair form, and the like). Pointing allows us to escape the circle of words.

The key role that examples play in our account accords well with the *kind* of term I take the ordinary word 'race' to be. I believe that 'race' belongs to the class of ordinary words whose application Tyler Burge characterizes as "indexically-mediated" and "perceptually backed."[26] Its meaning is fixed in part by reference to the perceptible examples to which it is applied. Its original home is in ordinary experience. It differs dramatically from specifically theoretical terms, such as 'phlogiston', which are born of a systematic theory and not indexically applied.[27] In this respect, it resembles a wide range of ordinary terms (Burge cites 'sofa' as an example) (*op. cit.*). Criticisms of the word 'race' based on the demand for precise criteria are misplaced.[28] Being *indexically mediated*, the word can be applied in the *absence* of precise criteria.

I should also note that possession of the ordinary concept of race does not require the ability to make subtle judgments about differences between racial groups. Nor does it require the ability to identify particular racial groups with any specified degree of accuracy. It certainly does not require the ability to assign each individual to one and only one racial group. What possession does require is the ability to identify some visible physical features as racial and to distinguish racial groups in a rough and ready way. 'Rough and ready' *should* be underscored. The requisite ability is very rudimentary.

(2) The concept of race is the concept of a group of human beings whose members are linked by a common ancestry.

Taken by itself, the first thesis might suggest that race is a matter of physical appearance *merely*. The second thesis corrects this misimpression.[29] It asserts that race is also a matter of ancestry: who one's parents are, who their parents are, and so forth. The link between race and ancestry is enshrined in the etymology of the English word 'race', which, the dictionaries tell us, derives from the Italian term 'razza' (meaning "race" and "lineage"). The German 'Rass' and French 'race' have the same derivation. A race is a lineage, a line. The infamous "one drop rule" ("one drop" of "black blood" makes an individual

[26] "Intellectual Norms and Foundations of Mind," this JOURNAL, LXXXIII, 12 (December 1986): 697–720, see pp. 703, 716. Burge himself does not discuss the word 'race'.

[27] Burge contrasts 'phlogiston' with ordinary words, p 716n. On race and 'phlogiston' see Montagu, "Concept," p. 19. See also Zack, p. 6.

[28] Cf. Appiah, "Race," p 42.

[29] My discussion of the ancestry condition draws upon Kitcher, especially pp 92–97.

black, whatever his or her outward appearance might be) is not a part of the ordinary concept of race. It does however reflect—albeit in a grotesque, distorted yet terribly vivid way—the idea that ancestry is central to race.

Now the ordinary concept of race does not itself specify the details of the nature of the ancestry that racial identity involves. Different conceptions of race will specify them in different ways. Nonetheless it is possible to extract from the concept a number of structural features. Thus, for example, an individual's race (membership in a particular racial group) is determined by the race (racial membership) of the individual's *immediate* ancestors: his or her parents. If both parents of an individual belong to one particular racial group *R*, that individual will belong to *R*, and belong to *R*—by virtue of his ancestry. If the parents of an individual belong to a particular racial group *R*, it is because *they* are the offspring of parents who belong to *R*.[50] Race is something inherited, a property transmitted from generation to generation.

Now the idea of a lineage brings with it the idea of a *founder*, a group of individuals who *found* the line.[51] Biologists refer to such groups as *founder populations*.[52] The line that the founders form is maintained through inbreeding. Endogamy is thus a structural characteristic of race. For a racial lineage to continue as the lineage it is, its members must for the most part select mates from within the racial group. I interpret the requirement of endogamy loosely.

The logical core of the ordinary concept of race does not exclude the possibility of a racial lineage tolerating a certain amount of interbreeding (consensual or otherwise). There is presumably a point at which exogamy leads to dissolution of the racial group. That point, however, is not specified by the concept.

[50] Whether an individual can belong to more than one race is a question which the ordinary concept of race does not answer. Particular *conceptions* of race may stipulate that each individual—or each "normal" individual—belongs to one and only one race as the ordinary conception is sometimes thought to do. But the ordinary concept of race itself contains no such proviso.

[51] See Kitcher, p. 93.

[52] Mayr defines a (local) population as a "community of interbreeding individuals at a given location"—*Populations, Species, and Evolution* (Cambridge: Harvard, 1963) p. 82. Zack follows Goodman in arguing that racial groups cannot be conceived of as populations "because races are non-evolving, ideal-types whereas populations constantly evolve" (p. 68). Goodman says "if races become populations, they are no longer races"—"Six Wrongs of Racial Science," in Curtis Stokes, Teresa Meléndez, and Genice Rhodes-Reed, eds., *Race in 21st Century America* (East Lansing: Michigan State UP, 2001), pp. 25–47, see p. 29. Their shared position presupposes that race must be conceived of essentialistically. I believe this is a mistake—one that rests on the common confusion of concept and conception—p. 439–40. I argue that the ordinary conception *is* and the ordinary concept of race *is not* essentialistic on pp. 449 and 453 respectively.

What makes a lineage *racial*? We can arrive at an answer by combining the first and second theses. A racial lineage consists of (a) a succession of human groups distinguished from other human groups by certain visible physical features and (b) a common founding group. The founders must (collectively) exhibit distinctive visible features of the relevant kind or possess alleles for such features. The same holds true of their offspring. If a group's members share the same visible physical features but do not descend from a common founding group, that group does not count as a race. If the members share a common founding group but do not resemble one another in their visible physical features, their group does not count as a race.

The third thesis holds that

(3) The concept of race is the concept of a group of human beings who originate from a distinctive geographic location.

The association between geography and racial group can be seen in the names the major writers on race in the seventeenth century assigned to racial groups. Thus, for example, Linneaus offered the labels 'Americanus', 'Europaeus', 'Asiaticus', and 'Afer' [African].[33] Buffon used the labels 'American', 'African', 'Chinese', 'Laplander', and 'European'.[34] The terms favored by Blumenbach are 'Caucasian' (from Mt. Caucasus), 'Monglian', 'Ethiopian', 'Asian', and 'Malay.'[35] The association is preserved on or below the surface of many names used for racial groups today. 'White' (black, Asian) is the name for the group whose aboriginal habitat is Europe (Sub-Saharan Africa, Asia).[36]

The lexical link between race and geographical area reflects the intuitive idea that racial groups have their origin in different geographical locations. If R_1 is a racial group, there is a specific geographical location G_1 uniquely associated with it that is distinct from some other geographical location G_2 uniquely associated with some other racial group R_2. The geographical location associated with a racial group consists of the "aboriginal" habitat of its ancestors. The aboriginal habitat of common-sense and conventional races is fixed by the location the ancestors of the members of those groups occupied

[33] *System Naturae* (1758) in *Race and the Enlightenment*, Emmanuel Chukwudi Eze, ed (Cambridge: Blackwell, 1997), pp. 10–14.

[34] *Histoire Naturelle Générale Et Particulière* (1769) *Natural History, General and Particular*, in Eze, ed, pp 15–28.

[35] *De generis humani varietate nativa*, Third Edition (1795) *On the Natural Varieties of Mankind*, in Eze, ed., pp 79–90.

[36] The term 'African-American' displays the lexical link of race and geography, on the surface while narrowing the extension of 'black'.

immediately prior to the advent of European oceangoing transport, which is to say around 1492.[37]

The archetypical example of the aboriginal habitat of racial groups is the continent.[38] Thus Africa is thought to constitute the aboriginal habitat of blacks, Europe the aboriginal habitat of whites, and Asia the aboriginal habitat of Asians. A racial group's distinctive visible features are the product of its ancestors' adaptation to the climate of its aboriginal habitat.[39] The visible physical features of race can thus be thought of as adaptive reactions to climatic variation. The fact that these features are features of this kind helps to explain why the visible physical features of race turn out to be the particular features they are.[40] What skin color, hair form, body shape, and the like all have in common is that they are all visible adaptive reactions to climatic variation.

The essential distinguishing characteristic of the aboriginal habitats of racial groups is geographical separation. This separation helps to explain how it was that different racial groups came to be exposed to different climatic conditions. Their habitats occupied different locations. The physical barriers that divided the different aboriginal habitats (mountains, deserts, oceans) prior to 1492 impeded interbreeding. Geographical separation resulted in (weak) reproductive isolation. The combination of this isolation and the differences in the climatic conditions of the different aboriginal habitats helps to explain how it was that the different racial groups came to have *different* visible physical features. The persistence of these differences after 1492—which is to say, after the development of technologies that made it possible to surmount the physical obstacles—is due in no small part to the fact that *social* barriers (lynching, law, custom, income) came to replace the physical ones in the interim.

[37] History books tell us that the Portuguese exploration of the African coast began in 1434. The dates 1450 and 1500 are also conventionally used to mark the relevant historical juncture.

[38] The idea of the continent as a naturally given geographical unit has come under fire in recent times. Continents, like races, are said to be "socially constructed"—see, for example, Martin W. Lewis and Karen E. Wigen, *The Myth of Continents* (Berkeley: California UP, 1997). It is of interest that the formation of the concept of continent and that of the (modern) concept of race were roughly coeval. One wonders if the two ideas are mutually entwined. Could it be that our idea of a *continent* (and hence *subcontinent*) derives in part from the idea of the *habitat of a racial group*? Could it be that the idea of a *racial group* gets part of its content from the idea of a group whose aboriginal home is a distinctive *continent*? Perhaps the concepts should be thought of as operating in tandem, each helping to fix the reference of the other.

[39] Cavalli-Sforza, pp. 10–13; Mayr, *Animal Species and Evolution*, pp. 645–46.

[40] The characterization of the visible features of race as adaptive reactions to climatic variation is not uncontested—see, for example, Diamond, p. 88; Zack, pp. 39–40, 56–57.

One additional point: a basic understanding of the logical core also requires a basic grasp of what it does *not* include. Thus, for example, it does not require that the founding group exhibit some kind of genetic "purity" (a high-degree of genetic homogeneity at the time of racial origination).[41] More specifically, it is compatible with the modern view that there are "no 'pure' human groups who have existed since the Creation as separate units."[42] Nor, again, does it require the existence of *sharp divisions* between different racial groups.[43] More precisely, it does not demand that each race R_1 possess some unique property P (visible or hidden) or set of properties S that it shares with no other race (R_2, R_3, R_4). A fortiori the ordinary concept of race does not require that each race R_1 possess some *fixed and immutable* unique property P (visible or hidden) or set of properties S that it shares with no other race.[44] It follows from this that the logical core does not require that races have *essences*. The ordinary concept of race is not essentialistic.

Being nonessentialist, the ordinary concept of race is nonracialist, racialism being a species of essentialism.[45] Its logical core does not hold that human beings are divided "by nature" into a hierarchy of races. Nor does it hold that each race is characterized by a fixed set of fundamental, "heritable," moral, intellectual, and cultural characteristics common and peculiar to it. Nor again does it hold that each race has an essence that explains why it has the visible physical features it does, why it has the moral, intellectual, and cultural characteristics it does, or why the two are correlated. It does not hold that each individual member of a race necessarily shares the "essential" characteristics of his or her race. It does not—and this point should be underscored—require *any* intrinsic connection between skin color and humanly important traits such as intelligence or moral character.[46]

But to say that the logical core of the ordinary concept of race is *nonessentialist* is not to say that it is "*antiessentialist*." The logical core

[41] I borrow this term from Kitcher, p. 98.

[42] Lewontin, p. 113.

[43] Montagu, "Concept," p. 18; James Shreeve, "Terms of Estrangement," *Discover* (November 1994): 60; Zack, pp. 43–44, 63.

[44] Goodman, p. 27; Zack, p. 68.

[45] I draw a distinction between 'racism' and 'racialism'. Very briefly, 'racialism' is the name of a theory or family of theories about the human species and its subdivisions. It works in the third person. 'Racism' by contrast is an affective and voluntative attitude toward individuals and groups. It works in the second person. Defining 'racism' as the view that there are fixed differences between the groups we call racial or as the view that there is a natural hierarchy among races has the disadvantage of making it difficult to discuss the view (which I call "racialist") in a noninflammatory way.

[46] Cf. Montague, "Concept," p. 24; Appiah, "Race," p. 7.

is compatible with both antiessentialist *and* essentialist conceptions of race. This is to be expected. The logical core "should" (in some suitable sense of 'should') be such that it can be used to articulate conceptions of race of both kinds. It is because essentialist and antiessentialist conceptions of race share a common subject matter fixed by the ordinary concept or race that they are *able* to contradict one another.

The logical core of the ordinary concept of race is not committed to any particular taxonomy. It does not purport to tell us how many races there are or whether race even has a determinate cardinality. Its resources are too restricted to settle these questions. To address any one of them (or to dismiss them all as wrongheaded), the core would have to be supplemented by a more determinate conception of race. The basic list of common-sense races (white, black, Asian) is fixed by the ordinary *conception* of race. The coherence of the ordinary concept of race therefore does not depend on the availability of any particular racial taxonomy.[47]

The logical core is compatible with the modern view that "the [scientific] term 'race' designates a group or population characterized by some concentrations, relative as to frequency and distribution, of hereditary particles (genes) or physical characters, which appear, fluctuate, and often disappear in the course of time by reason of geographic and/or cultural isolation."[48] It does not require that the characters of members of racial groups be *concordant* (co-vary). And it is compatible with the modern view that the alleles of members of different races (including dispositions to phenotype) differ only in frequency (*ibid.*) and that *intra*racial diversity (within-population genetic variation) is far more pronounced than *extra*racial diversity (among-populations genetic variation).[49] The logical core is also compatible with the generally recognized fact that there is no single diagnostic trait of racial membership. Taken together, these considerations suggest that to say that "in the light of modern genetics" the logical core of the ordinary concept of race is "*utterly* erroneous and meaningless," is something of an exaggeration.[50]

[47] Cf. Diamond, p. 84; Zack, pp. 1, 41–57, 69.

[48] UNESCO, "First Statement on Race," in Montagu, *Statement on Race* (New York: Oxford, 1972), pp. 1–278, p. 46.

[49] In a recent study Noah A. Rosenberg et al., found that a whopping 93–95% of the average proportion of human genetic variation is due to differences between individuals within the same population and that only a miniscule 3–5% is due to differences among populations— "Genetic Structure of Human Populations," *Science*, CCXCVIII (December 2002). 2381–85, see p. 2381. See also the earlier Lewontin, p. 123

[50] Montagu, "Species," p. 3 (emphasis added).

One of the most striking results of our account of the logical core of the ordinary concept of race is that race turns out to be relatively unimportant. As represented by the logical core, race is a matter of skin color, shape, ancestry, and aboriginal habitat—*and nothing more.* Now the ordinary *concept* of race can be used to articulate *conceptions* of race that make race important (for example, by linking skin color to moral character). But the logical core itself does not make it important. It is therefore an error to think that an account of the concept of race *must* explain its intrinsic importance if it is to be adequate.[51] Intrinsic importance is precisely what the concept lacks.

But *if* the logical core is combined with other things, it can *become* important. Thus, for example, if, in a particular society, being black (in the sense of skin color and ancestry) is a risk factor for hypertension (asthma, diabetes, heart failure, the list goes on) then being black-in-that-society is a risk factor for these diseases and hence important from the standpoint of health and medicine.[52] If, in a particular society, the *color* black is understood to mean "[d]eeply stained with dirt; soiled, dirty, foul...having dark or deadly purposes, malignant; pertaining to or involving death, deadly, baneful, disastrous, sinister...Foul, iniquitous, atrocious, horrible, wicked;...Indicating disgrace, censure, liability to punishment, and so forth,"[53] as was the case in England before the sixteenth century, then being black (having skin of a color called black) will be imbued with intense symbolic significance in that society. Again, if race (in the sense of color and ancestry) is used to assign individuals to specific social positions or correlates with their individual life prospects, then what one's color is and who one's ancestors are will be of tremendous social importance.

II.B. We are now in a position to turn to the *racialist development* of the concept of race. The racialist development fixes the ordinary *conception* of race. It is best thought of as a logical construction based on the logical core. The construction starts with the three theses specifying the ordinary concept's logical core:

(1) The concept of race is the concept of a group of human beings distinguished from other groups of human beings by visible physical features of the relevant kind.

(2) The concept of race is the concept of a group of human beings whose members are linked by a common ancestry.

[51] Cf. Appiah, "Race," p. 74.

[52] On this topic, see Michael Root, "The Use of Race in Medicine as a Proxy for Genetic Differences," (forthcoming in *Philosophy of Science, Proceedings*).

[53] *Oxford English Dictionary*, cited in Winthrop D Jordan, *White over Black* (Williamsburg: North Carolina UP, 1968), p. 7

(3) The concept of race is the concept of a group of human beings who originate from a distinctive geographic location.

and then appends to these six additional theses, drawn from the characterization of racialism considered earlier.

The concept of race is the concept of

(4) a natural division of the human species into a hierarchy of groups that satisfy the conditions specified in (1)–(3).
(5) a group of human beings satisfying the conditions specified in (1)–(3) which is characterized by a fixed set of fundamental, "heritable," physical, moral, intellectual, and cultural characteristics common and peculiar to it.
(6) a group of human beings satisfying the conditions specified in (1)–(3) whose distinctive visible physical features are correlated with the moral, intellectual, and cultural characteristics that are common and peculiar to it.
(7) a group satisfying the conditions specified in (1)–(3) that possesses an "essence" which explains why it is that the group has the distinctive visible features it does, why it is that the group has the particular moral, intellectual, and cultural characteristics it does, and the correlation between the two.
(8) a group of human beings satisfying the conditions specified in (1)–(3) whose members necessarily share its "essential" characteristics.
(9) a group of human being satisfying the conditions specified in (1)–(3) whose essential characteristics constitute the essence of its members.

Looking at this list, we see that the racialist development of race is the product of adding theses (4)–(9) to (1)–(3). Theses (4)–(9) represent the specifically *racialist* components of this conception. It is evident that the logical core, (1)–(3), is prior to and independent of the racialist development (4)–(9). It is also plain that the racialist development is something *added on* to the logical core. If the ordinary concept of race is no ordinary ordinary concept, it is because it has been subject to the racialist development.

Now from our standpoint, the step from the logical core (3) to racialist development (4) seems huge. *If*, however, one were (i) already under the sway of the modern taxonomic project (ii) still in the thrall of Aristotelian (or Platonic) essentialism, (iii) chauvinistic about one's own racial group, (iv) ethnocentric, and (v) struck by the differences in appearance of the aboriginal inhabitants of other continents, all of which is to say that *if* one had the mindset characteristic of eighteenth and nineteenth centuries—*then* the step from logical core to racialist development would appear to be short. The racialist development would present itself as a natural extension of the logical core.

Or it would if the logical core and racialist development were distinguished—which, in the eighteenth and nineteenth century, they were not.

When the logical core first entered the historical scene, it was *already articulated* by the racialist development. The step from logical core to racialist development is logical not temporal. But the fact that this logical step *was* taken is *historically* contingent. Had world history gone differently—very differently perhaps—the logical core could have appeared without the racialist development. Nothing in the logical core *necessitates* this step. But in point of fact, things did *not* go differently. And because of this, the logical core and racialist development *appear* to be a unity.

Let us compare the racialist development with the logical core. Unlike the logical core, the racialist development *is* inherently "essentialist." It does require a sharp distinction between different races. In contrast to the logical core, it is hierarchical. It represents the social hierarchy of race as "natural." The racialist development also differs from the logical core in that it conflicts with current biological thinking; in contradistinction to the logical core, it *is* essentialistic. There is no room in the modern biology (in the modern conception of species) that species have distinct essences which distinguish them from other species.[54] A fortiori there is no room in biology for essences at the *infraspecific* level.[55]

Another point of contrast is that the racialist development represents race as being intrinsically important. It says that the visual physical features of race *are* correlated with humanly important characteristics. To be of a particular race is to have a particular set of moral, intellectual, and cultural aptitudes and tendencies. The racialist conception represents the visible physical features of race as reliable markers of humanly important differences.

The racialist development and the logical core differ in their normative force. The logical core does not provide a rationale for racism, slavery, colonialism, or genocide. The racialist development does. It motivates the step from (a) representing another group as racially different to (b) taking these differences to be humanly important, to (c) regarding the other group as inferior, and (d) making it the object of hatred and contempt, to (e) imposing upon it involuntary

[54] Elliott Sober is especially clear on this point: "No phenotypic characteristics can be postulated as a species essence; the norm of reaction for each genotype shows that it is arbitrary to single out as privileged one phenotype as opposed to any other. Similar considerations show that no genotypic characteristics can be postulated as a species essence"—"Evolution, Population Thinking, Essentialism," in *From a Biological Point of View* (New York: Cambridge, 1994), pp. 201–32, here p. 225.

[55] This does not entail that the ordinary concept of race is *biologically significant*.

servitude or (f) colonial rule, or (g) attempting the liquidation of all its members—a sequence of steps historically all too familiar.

The racialist development is inherently malignant. The logical core is not. The harm associated with the concept of race is due to its racialist development and not its logical core. We should, however, be wary of concluding that the logical core is benign. For it is extremely easy to slide *from* it *to* the racialist development. Ordinarily the two go together. It is only by force of abstraction that they can be separated. Considered from this angle, the logical core is anything but benign.

III

I want to close by considering some of the implications of the present account for eliminativism about the concept of race. Eliminativism is the view that the concept of race ought to be "eliminated" (withdrawn from use in theory and talk) on the grounds that it provides justification and support for racism. This position is based on the premise that the concept of race is inherently racialist. I have argued against this premise. What *is* racialist is the conception of race built into the racialist development of the concept. This raises the possibility that eliminativism may have misidentified its target. Perhaps its real or proper object is racialism or the *racialist development* of the concept of race—not the ordinary concept of race itself.

This is important because the ordinary concept of race is rather difficult to do without. For one thing, it is needed for an analysis of *racism*. Racism is (roughly) hostility to individuals on the basis of their *race*. That is to say, on the basis of their (alleged) membership in a *racial* group: a group whose members exhibit a set of distinctive visible characteristics associated with their ancestry and aboriginal home. It is not easy—perhaps not even possible—to make sense of the object of racism and hence not possible to make sense of racism itself without employing a concept of race. We therefore cannot jettison the concept of race without at the same time jettisoning an indispensable tool in the struggle against racism: the concept of racism itself.

Another reason the ordinary concept of race is difficult to do without is this. The concept captures a striking, if superficial, feature of empirical reality. Human beings exhibit differences of shape and color which are associated with differences in ancestry and aboriginal habitat. *These* differences are real. In saying this, we say nothing about any alleged moral, intellectual, or cultural differences that are alleged to be correlated with the former. And it is undeniable that these visible features are striking—from a *human* point of view. The proof being that human beings *find them* striking. The idea that the salience of these visible physical differences is *fully* explained by the existence of racism, I frankly find unbelievable. Consider a world without racism.

Differences of shape and color would, no doubt, have a different *significance* in such a world. For one thing they would be *much much* less important. But it seems to me likely that they would still be striking—*physically* striking—in a world free of racism. And, striking or not, these differences are *differences*. The concept of race is needed to represent these differences and capture their specific character.

The case for eliminativism rests in part on the simple idea that without *race* there can be no *racism*. No race, no racism. In the absence of differences of shape and color, so the argument goes, racism would have no toehold in reality. The latter point is no doubt correct. The problem is that racism *does* have a toehold in reality. Human beings *do* differ in shape and color in ways that are connected to ancestry and aboriginal habitat. Pretending that what is the case is *not* the case cannot be an effective strategy for combating racism. Nor can denying that these differences are racial. Their existence is not, however, a cause for despair. Science tells us that they have no intrinsic significance and are not correlated with traits that are humanly significant. There is a gap between race and racialism. Racism's toehold in reality is just that, a toehold—and nothing more.

Our account of the ordinary conception of race makes it possible to see that the word 'race' can be used to express the logical core of the concept without drawing upon its racialist development. It shows the possibility of using the concept of race without thinking racist thoughts. Given the difficulty of jettisoning the concept of race, this is an important result.

It would, however, be a mistake to conclude that the ordinary concept of race is unproblematic. The logical core and racialist development of the ordinary concept of race, although conceptually separable, are intimately connected. The precise nature of this connection is complex and not entirely clear. But whatever the exact nature of this connection (psychological, cultural, semiotic), it goes very deep. The existence of this connection makes it unrealistic to suppose that the word 'race' can ever be wholly liberated from its racialist overtones and connotations. The concept is indelibly stained by its history and associations. To use the word 'race' is to run a series of risks: the risk of being misunderstood, of drawing unwittingly upon the concept's racialist development, and perhaps even of inadvertently providing racism with some modicum of support. These risks, however, cannot be avoided since the fact of racism means we must talk of race. The lesson to be drawn here is that we need to learn to navigate with a concept that is indelibly stained and ineluctably hazardous. This burden, itself a legacy of racism, comes with the territory.

MICHAEL O. HARDIMON
University of California/San Diego

ON THE NEW BIOLOGY OF RACE*

The notion that race has some biological basis has been widely criticized, by both biologists and philosophers. Indeed, the view that race is no more scientifically real than witchcraft is so influential that many who want to argue that race is real divert to understandings of race and reality according to which race is real as a social, rather than natural, kind.[1] Against this trend, however, Robin Andreasen and Philip Kitcher have recently argued for an improved biology of race.[2] The improvements over past biological accounts of race are two-fold. First, the new biology of race avoids the racism of prior biological accounts of race, which often attributed intrinsic significance to racial phenotypic traits or tied intellectual, aesthetic, cultural, and moral potential to those traits. Indeed, both Andreasen and Kitcher, while trying to make biological sense of race, reject the conflict and social division that has surrounded race for so long. Second, the new biology of race actually includes sound scientific research. Briefly, the key idea to this new biology of race is that while perhaps there is no "race gene" or set of necessary and jointly sufficient phenotypic features that can be attributed to each race, races can be understood as *breeding populations*. Here I want to question the viability of this approach.

I

On Andreasen's "cladistic approach," we need not categorize humans according to inherent essences, geography, or conventionally established similarities, for there is another option for a viable biological understanding of race: we can classify via genealogy. A cladistic classification can be represented in a phylogenetic tree, as in Figure 1.[3]

* Thanks to David Eng for helpful comments on earlier drafts.

[1] See, for example, Michael Root, "How We Divide the World," *Philosophy of Science*, LXVII, Supplementary Volume (2000): S628–39; Ronald R. Sundstrom, "Race as a Human Kind," *Philosophy and Social Criticism*, XXVIII (2002): 91–115; Sundstrom, "'Racial Nominalism'," *Journal of Social Philosophy*, XXXIII (2002): 193–210; and Paul C. Taylor, "Appiah's Uncompleted Argument: W.E.B. Du Bois and the Reality of Race," *Social Theory and Practice*, XXVI (2000): 103–28

[2] Andreasen, "A New Perspective on the Race Debate," *British Journal for the Philosophy of Science*, XLIX (1998): 199–225; Andreasen, "Race: Biological Reality or Social Construct?" *Philosophy of Science*, LXVII Supplementary Volume (2000): S653–66; Kitcher, "Race, Ethnicity, Biology, Culture," in Leonard Harris, ed., *Racism* (Amherst, NY: Humanity, 1999), pp. 87–120.

[3] Andreasen, "A New Perspective," p. 207.

0022-362X/03/0009/456–74 © 2003 The Journal of Philosophy, Inc.

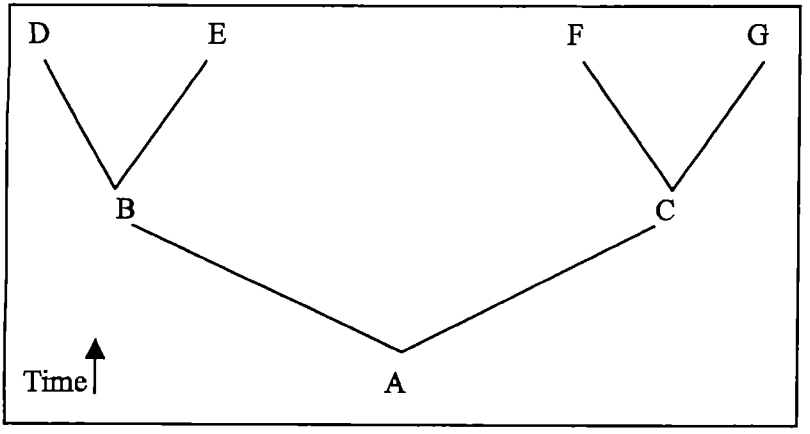


Figure 1

Each letter in Figure 1 represents a breeding population, that is, a population that is largely reproductively isolated from other populations. Using this model, Andreasen constructs a phylogenetic tree, and hence identifies distinct breeding populations in terms of the relative closeness of different human populations through genetic distance, "a measure of the difference in gene frequencies between two breeding populations" (*ibid.*, p. 210).[4] Drawing from biological data, Andreasen informs us that Africans and non-Africans have the furthest genetic distance. The second split separates Pacific and South East Asians from other non-Africans, and eventually splits occur between Pacific Islanders and South East Asians, North Eurasians and Caucasians, and finally within North Eurasians. On the basis of these data, Andreasen borrows the tree in Figure 2 from Luigi Luca Cavalli-Sforza.[5]

Since biologists can construct such a "family tree," Andreasen concludes the following: "It means that it is possible to give a biologically objective definition of race. Races are monophyletic groups; they are ancestor-descendant sequences of breeding populations, or groups of such sequences, that share a common origin."[6]

[4] It is worth noting that reproductive isolation does not exclusively mean geographic isolation; other mechanisms can foster isolation, as we will see in our discussion of Kitcher. In addition, the isolation does not have to be so strong that there is zero interpopulation reproduction. It only entails that there is enough difference between intrapopulation reproduction and interpopulation reproduction to limit gene flow between populations, thereby resulting in genetic distance.

[5] Andreasen, "A New Perspective," p. 212, and "Race," p. S660, Cavalli-Sforza, "Genes, Peoples, and Languages," *Scientific American*, CCLXV, 5 (1991): 104–10.

[6] Andreasen, "A New Perspective," p. 214.

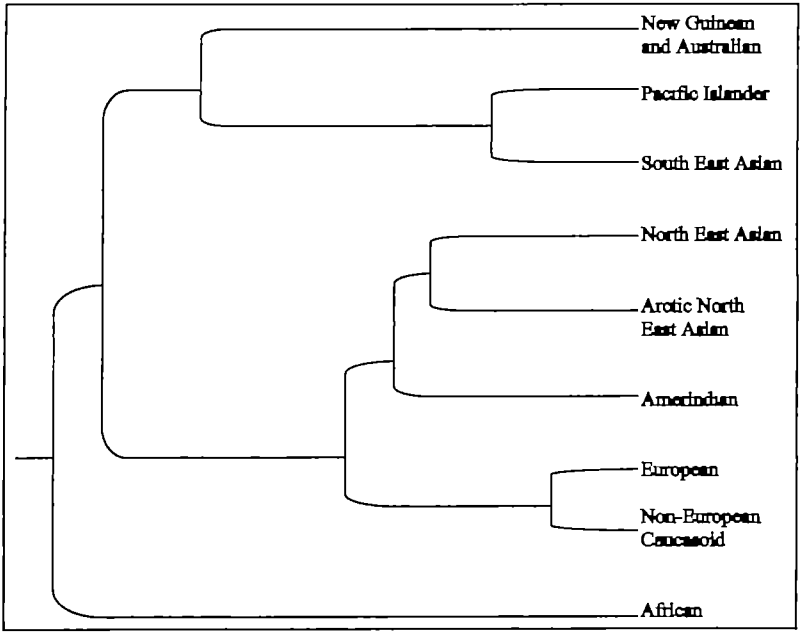


Figure 2

This account has several virtues.[7] First, and most obviously, it seems to give a biological backing to our race talk. Second, since it hinges on reproductive isolation, the cladistic approach reveals the dynamic nature of race. That is, insofar as reproductive isolation has increasingly eroded since the European "discovery" and colonization period, races have been slowly burning out of existence for the last 500 or so years. Thus, we have racial ancestries, even if there are no current biological races. (As we will see, the view that races are disappearing marks a significant point of distinction between Andreasen and Kitcher.) Finally, the cladistic model carries no racist baggage, unlike so many preceding biological notions of race.

II

I have no quarrel with the second and third virtues just mentioned,[8] but I think we ought to take a closer look at the first point—that the cladistic approach affords race some biological reality. Consider first

[7] Andreasen, "A New Perspective," pp. 215–17, and "Race," p. S664.

[8] Nor with the science behind Andreasen's approach—she considers some possible objections on that front in "A New Perspective."

that the folk notion of race does not normally contain the nine races identified in Figure 2. As others have noted, it is difficult to determine exactly what races the folk concept of race includes: some speak of "three major races"—African, Asian, and Caucasian; others consider Latinos, or Native Americans, to constitute a race. (Just in terms of numbers, I think it is safe to say that currently it is rare to hear a folk notion of race that involves more than four or five races).[9]

Accordingly, it seems that the nine races in Figure 2 do not correspond extensionally with folk notions of race, and Andreasen is quick to agree, with particular mention of the Asian varieties in her schema: "the folk category 'Asian' is not a cladistic race.... North East Asians are more closely related to Amerindians and to Caucasians than they are to South East Asians. Similarly, South East Asians are more closely related to Australians than to North East Asians." Andreasen takes this discontinuity with the folk notion of 'race' to be nonproblematic, when she continues, "This conclusion is interesting because it illustrates that the existence of biological races does not depend upon our folk taxonomy being right."[10]

But this exposes what I take to be a central flaw of the cladistic approach. That is, Andreasen has found a way of carving our ancestors into breeding populations, but these populations are not what we call 'races'. In addition to the extensional differences already noted, consider the intension of 'race'. Intensionally, of course, 'race' can mean (and has meant, over the years) a number of different things.[11]

[9] As is often pointed out, such folk categorizations often inconsistently offer groupings that overlap race, ethnicity, and national origin. The history of intellectuals theorizing about race—as opposed to folk categorization—offers an extremely varied set of lists of the races, which differ not only on how many races there are, but also on which races there are (none of which seems to match Andreasen's list). To mention just a few examples, Bernier lists four or five (he is noncommittal about whether Native Americans constitute a distinct race); Voltaire offers seven; Kant offers four or five, depending on the essay; and Du Bois eight. As Robert Bernasconi and Tommy Lott note, by the end of the nineteenth century, the number of races "grew from four or five to fifty or even eighty," except in the U.S., which sought to condense everyone of European descent into one race, to the exclusion of blacks, Asians, and Native Americans in particular—"Introduction," in Bernasconi and Lott, ed., *The Idea of Race* (Indianapolis: Hackett, 2000), p. x. As the question posed here is whether there is any biological referent to what we, especially in the U.S., currently identify as races, I will be concerned with whether Andreasen and Kitcher can—or even need to—account for a biological basis of the current *folk* concept of race.

[10] Andreasen, "A New Perspective," p. 212–13; cf. Andreasen, "Race," p. S664.

[11] For two detailed analyses of what 'race' means and has meant, see K. Anthony Appiah, "Race, Culture, Identity: Misunderstood Connections," in Appiah and Amy Gutmann, *Color Conscious: The Political Morality of Race* (Princeton: University Press, 1996), and David Theo Goldberg, *Racist Culture: Philosophy and the Politics of Meaning* (Cambridge: Blackwell, 1993), chapter 4. Here, I only make some intuitive observations about what people mean or have meant by 'race.' I take it, though these observations are made from the armchair, so to speak, it is more or less obvious that these have been, at one

It might mean something as putatively benign as groupings based on pigmentation, for instance. Andreasen, however, holds that "Individuals are members of a cladistic race if and only if they belong to breeding populations that share a common origin. This will be true regardless of how closely they resemble each other."[12] Accordingly, her cladistic classification does not match up with the folk concept of race that centers on pigmentation. Presumably, for example, we would at least struggle to reconcile this folk concept with the idea that "South East Asians are more closely related to Australians than to North East Asians." None of this means that the ordinary notion of 'race' is coherent; the point is simply that the cladistic approach does not provide biological backing for *it*.

Or 'race' might mean something else. It might refer to a cluster of phenotypic features (in addition to skin color), but, since physical resemblance is irrelevant on Andreasen's cladistic approach, that approach seems incapable of matching a phenotypic-cluster folk classification. Or, 'race' might be taken in the way (some) racists mean it, when they attach intellectual, moral, or aesthetic characteristics to phenotypic features in a hierarchical fashion. Or, it might be a concept that attaches some such characteristics to phenotypic "markers," but in a nonhierarchical way, as we find in W.E.B. Du Bois.[13] The point here is not that one of these understandings is better than the others (and, to be sure, each has its problems); again, the point is that Andreasen's cladistic approach does not map onto *any* of the more dominant folk conceptions of 'race', insofar as those conceptions are about more than genealogy.

Surely, however, ancestry—which is at the heart of the cladistic approach—does play a large role in many conventional understandings of race. (This is one horn of Andreasen's two-pronged response

---

time or another, conventional notions of 'race'. Andreasen offers some overlapping characterizations of common-sense ideas of race in "Race," p. S663. See Luther Wright, Jr., "Who's Black, Who's White, and Who Cares: Reconceptualizing the United States's Definition of Race and Racial Classifications," *Vanderbilt Law Review*, XLVIII (1995). 513–69, for summary and analysis of legal definitions of race in the U.S. Two relevant results can be found there: when not conflated with ethnicity and national origin, the legal definition of 'race' usually boils down to either overt physical traits or descent. (And, since on the second criterion one is of race R when both of one's parents are of R, presumably parental racial classification at some point must be defined by some non-genealogical criterion, most likely physical traits.) Finally, see Charles Hirschman, Richard Alba, and Reynolds Farley, "The Meaning and Measurement of Race in the U.S. Census: Glimpses into the Future," *Demography*, XXXVII (2000): 381–93, for how U.S. citizens self-identify in census reporting.

[12] Andreasen, "Race," p. S664.

[13] "The Conservation of Races," reprinted in Du Bois, *The Souls of Black Folk*, David W. Blight and Robert Gooding-Williams, eds. (Boston: Bedford, 1997).

to the type of objection offered here.[14]) In the United States, at any rate, the "one-drop rule" has had a crucial role in our system of racial classification, so that persons of mixed black-white ancestry often get labeled as (and identify as) black, morphological indicators notwithstanding. As has been pointed out, of course, the one-drop rule regarding blackness is problematic on several levels, including being inconsistent with U.S. policy regarding Native Americans.[15] The cladistic approach need not rely on the one-drop rule, however, for there is a more general ancestral component to the common-sense notion of race: a person is of race *R* if and only if her parents are both members of *R*.

Yet even this general genealogical element in common-sense notions of race does not match the cladistic approach. First, while ancestry is often *part* of the folk meaning of race, people frequently mean more than ancestry when they use racial discourse; skin color, for example, seems like a central—and inextricable—part of the folk meaning of 'race'. Indeed, race is sometimes thought to include even more than skin color. Again, there are overt racists, who adhere to hierarchical racial essences, as well as those, like Du Bois, who seem to hold that the races, while not hierarchically ordered, still have essential characteristics beyond the phenotypic ones.

Second, whatever else conventional notions of race are intended to mean, they all seem to include the idea that races still exist. This, in turn, entails that the folk concept of 'race' at least means more than isolated reproductive groups that are vanishing (or have already vanished). Andreasen, by contrast, holds that because interracial reproduction eliminates isolated breeding populations, races are ceasing to exist. Thus, it seems that the viability of Andreasen's cladistic approach, with its emphasis on ancestral isolated reproductive groups and dismissal of the centrality of phenotypic traits, does not entail that race—in its common-sense meanings—is real.[16]

[14] Andreasen, "Race," p. S665.

[15] See, for instance, Naomi Zack, *Race and Mixed Race* (Philadelphia: Temple, 1998) and M. Annette Jaimes, "Some Kind of Indian," in Zack, ed., *American Mixed Race: The Culture of Microdiversity* (Lanham, MD: Rowman and Littlefield, 1995) pp. 188–55.

[16] In their analyses of census-style self reporting on the 1996 Racial and Ethnic Targeted Test, Hirschman et al., "Meaning and Measurement," note that American folk classifications might be more productively captured in terms of origin, rather than race, but these origins are importantly different from Andreasen's ancestral breeding populations. The origins identified in Hirschman et al. include more recent origins like 'Hispanic', and national origins like 'Ecuadoran', which do not map on to Andreasen's ancestral breeding populations (intensionally or extensionally).

III

This brings us to the methodological question that is at the heart of Andreasen's theory, for she agrees that she has not shown that race *in its usual meanings* is real: "Questions about biological classification can be about ordinary language classifications, or they can be about scientific classifications. For example, the question 'is there a biologically objective way to define race' could be asking whether biology vindicates our common-sense notions of race. Alternatively, it could be asking whether there are *any* biologically objective ways to divide humans into races."[17] As Andreasen sees it, the theory proposed by those who think that race is socially constructed rather than real seeks to answer the first question—about common-sense notions of race. The cladistic approach, not inconsistent with the constructionist view, seeks to find a biological notion of race, whether or not it matches up with common sense.

This agenda reveals the core question: How revisionist can one be about the meaning of 'race' and still call it 'race'?[18] For instance, one might argue that because biologists can (to a large extent, though not entirely) divide the human species into two groups, namely those with XX chromosomes and those with XY, there is a biological notion of race, with two races, female and male. The right response to such an argument, I think, is that while it is true that this is one way of dividing up people biologically, it does not converge with what either intellectuals in the race debate or those who employ common sense mean by 'race'. Therefore, the argument has not established that races are biologically real. On the other hand, consider an approach that gave biological backing to race, but only required comparatively minor revisions to the common-sense notion of 'race' (for example, it required putting a group under the racial category *R* that previously was not thought of as *R*, but which contained members who were phenotypically similar). Perhaps we would concede that, indeed, races are biologically real and that we should revise our notion of 'race' in this minimal way.

Thus, while some minimal revision to the meaning of 'race' (as for all definitions, of course) is allowable in the search for biological backing for race, we must stay fairly close to the vest, or we risk not talking about race at all. The question, again, is: How much revision

[17] Andreasen, "A New Perspective," p. 218.

[18] While Andreasen is working independently of common-sense notions of race, she is also engaging in dialogue with those in "the race debate," as indicated by the title of her paper, "A New Perspective on the Race Debate." As such, it seems all parties ought to be in the same neighborhood, more or less.

is allowable? I cannot offer a good answer to that question here, but I do not think it is necessary. (Although, in section IV, I will suggest a limiting condition on concept revision.) For it seems to me that the burden is on the revisionist to show that her revisions are warranted. We need an argument from Andreasen that we should still call her breeding populations 'races', even though, first, her nine populations do not correspond extensionally to what we usually identify as races, and, second, her concept of 'race' as breeding population does not agree intensionally with the folk concept, insofar as we normally mean something beyond mere isolated reproductive groups, such as groups demarcated by skin color.

Andreasen does offer a response to this concern (this is the other horn of the two-pronged response): "One can find in the history of science many instances to support the idea that the objectivity of a [natural] kind is not undermined by the fact that ordinary people have mistaken beliefs about its nature."[19] Here she cites examples such as whales. Common-sense belief tells us (or once told us, anyway) that whales are fish. However, science classifies whales as mammals. This disagreement between science and common sense does not mean that whales do not exist; rather, it merely means that common sense is wrong. The upshot is that we can replace common-sense concepts with scientific ones when common sense is mistaken.

I think, however, that the analogy between whales and race is tenuous. In the disagreement over the status of whales, the scientist and the layperson can point to a thing they mutually agree is called 'whale', and the scientist can explain why it is more naturally lumped together with mammals than with fish, in terms of common properties like warm-bloodedness. This is a disagreement over how to classify one anomalous species in an otherwise fixed classification schema.

In Andreasen's account of race, however, there is a wholesale reshuffling of the classification schema itself. If, for instance, Andreasen and a layperson were to pick out a person of a certain genealogy and phenotype, where the layperson would classify that person as (say) 'Asian', and Andreasen would classify her instead as 'North East Asian', there would be a much different ensuing dialogue than that in the case of whales. Since Andreasen's classification schema puts the person in question closer to Caucasians than to South East Asians (while South East Asians are closer to Australians than to that person), it would soon become clear that the disagreement is not primarily about where to put this person in a fixed schema of classification, as it was in the case

[19] Andreasen, "Race," p. S662; cf. S665.

of labeling whales either fish or mammals. Rather, the disagreement is over the classification schema itself, and, consequently, over each schema's underlying definition of race—one focuses on descent and breeding group, while the other focuses on descent and phenotype. This disagreement accounts for each party using a different label to refer to the same person, in contrast to the case of whales, where both parties can agree on one term—'whale'—to refer to the object picked out by that term.

At this point the layperson could reasonably assert that now the disagreement is not about how to classify the person in question; rather, it is about what classification schema and definition of race to adopt. And, if our layperson were informed that the common-sense understanding of race, centered on phenotype, is biologically unfounded, it would not be unreasonable for the layperson to reply that perhaps this simply means that there are no races. In this respect, the disagreement is very different from the whale case, where the layperson would be unreasonable to claim that there are no whales. In short, there is a stand-off here that was not present in the whale case, regarding which classification schema and definition of 'race' to choose. Reclassifying anomalous cases of misclassification (like whales) is not analogous to making wholesale changes in the classificatory system.

Accordingly, we need a further argument for revising our concept of race; saying that if we revised our concept of 'race', then it would be more similar to the case of whales than to some nonexistent kind like witches, is not itself a ground for revising our concept of race. Rather, one must argue for revising the concept of race as part of showing that race is more like whales than witchcraft. Only then, if we accept that prior argument, can we explore other conceptions of race.

As a final note, the following claims cannot fill in the missing argument: common-sense races are not biologically real, whereas cladistic races are; therefore, we should replace the folk notion with the cladistic notion. Such an argument begs the question of whether a biologically real notion of 'race' that is minimally related to the common-sense notion is preferable to a conventional, but biologically nonexistent, notion of 'race'. This, I take it, is one of the more crucial questions in the race debate, and any answer must argue for, rather than simply stipulate, one side over the other. Below (section VI), I will further examine the question of replacing or significantly revising common-sense racial discourse.

IV

The above argument against Andreasen's account of race relies on four key premises: (1) that account's nine races present a system of racial classification that is substantially different from folk racial classifications; this extensional difference is based on the facts that (2) the cladistic model's reproductively isolated groups are disappearing (and so its eroding ancestral element does not match up with the folk notion's persisting ancestral element) and (3) reproductively isolated groups do not map onto the phenotypical groupings that seem essential to the folk concept of 'race'; finally, (4) scientists are not themselves the arbiters of the meaning of 'race'. The first three points are about the meaning of race; the fourth generates a meta-question about how meaning gets settled in the first place. What, then, if there were a viable classification of human races that could disrupt one or more of these premises? Indeed, Kitcher's theory of race seems to bypass the objections in (1)–(3). In this section, I want to examine premise (4); I take up Kitcher's alternative in the next section.

The foregoing suggests that treating races as reproductively isolated breeding populations is too different from the common-sense notion of 'race' to provide an adequate biological account of race. To then conclude, however, that there is no biological basis of race might seem to presuppose a blanket premise that everyday folk, rather than professional scientists, have the authority to determine the meanings of purported natural kind terms. As a general principle, however, this presupposition is not so easily defended, particularly if we adopt the causal theory of reference as found in the work of Saul Kripke and Hilary Putnam.

In Appiah's analysis of the meaning of 'race', for instance, he adopts a method of "semantic deference" that follows Putnam's "linguistic division of labor," a justification for nonspecialists' engaging in discourse that uses terms with meanings that those nonspecialists cannot identify. The only way that such discourse is legitimate is if specialists can identify the meanings of those terms so that they may be freely used in folk discourse.[20] And, if we follow Putnam's Twin Earth thought experiment, we find that water, necessarily, means H_2O, even if nonspecialists point to a watery substance with a chemical composition of XYZ and call that 'water', in part because of the authority vested in specialists. Why, then, should it be a strike against Andreasen's model of biological racial realism if nonspecialists point to

[20] Appiah, "Race, Culture, Identity," p 41.

a set of entities (for example, the "three major races") and call those 'races' in a way that is different from what the cladistic approach would suggest? Thus one rejoinder to my objection to Andreasen might be that since the experts, rather than common sense, determine the meaning of natural kind terms, it is irrelevant whether the biological data can provide a referent for the folk concept of race.

As I see it, there are two problems with this rejoinder. The first is that semantic deference requires an expert definition of the term in question, and this requires both a defined set of experts and a (near) consensus on what the expert definition of the term is. Chemists, for example, can provide a unified expert judgment that water is composed of H_2O. In the case of race, presumably the experts include biologists and philosophers (among other parties, such as physical anthropologists). However, there is severe disagreement among those experts when it comes to race, and that disagreement exists on two levels. On one level, there is disagreement over which entities are supposed to be identified as races in the first place. As we have seen, Andreasen identifies the nine populations in Figure 2. At the same time, however, Appiah's own examination of the concept of 'race' identifies the "three major races" of African, Asian, and Caucasian. Kitcher also identifies those as the three major races.[21]

On another level, even if that first question could be settled, there is widespread disagreement as to whether there is any biological reality to those racial classifications, and, if so, what the underlying biological referent is. For example, while Kitcher argues for the biological reality of the three major races, Appiah disagrees. Also consider Andreasen's use of her main source of biological data, Cavalli-Sforza: while Andreasen takes those data to generate a plausible story about the biological reality of race, her source disagrees.[22] In this respect, then, the expert consensus that water is H_2O is very different from the ability of experts to fix the meaning (both extensional and intensional) and ontological status of race—unlike water, there simply is no decisive expert opinion on the nature of race.

A second, related, point from above is pertinent here as well. We give preference to the views of experts over common sense only under certain limiting conditions. For instance, we prefer the specialist definition of water as H_2O, even when nonspecialists identify both H_2O and XYZ as 'water', only when there is reasonable overlap between the

[21] Appiah, "Race, Culture, Identity," p. 77; Kitcher, p. 87. Cf. notes 9 and 11 above.

[22] Cavalli-Sforza, P. Menozzi, and A. Piazza, *The History and Geography of Human Genes* (Princeton: University Press, 1994). For Andreasen's take on this disagreement, see "A New Perspective," p. 213.

gross physical substances identified by specialists and nonspecialists concerning the object to be defined. To change the example, suppose chemists pointed to a substance *S*, with the chemical compound NaCl, and told nonspecialists that chemical analysis reveals that water is NaCl, despite the mass of nonspecialists calling substance *W* (with the chemical composition H_2O) 'water'. Nonspecialists would rightly respond that NaCl is called 'salt', not 'water'. This is the limiting condition on semantic deference: meanings of folk terms are determined from the ground up by folk usage, rather than from above by specialists, to the extent that technical categories have to overlap reasonably with the folk categories themselves (like the relationship between water and watery substances, but unlike that between salt and water). If there is no scientific backing to some given folk category, and if there is no reasonably overlapping technical category that does have scientific backing, then that is when we determine that there simply are no things of that kind (for example, witches). This point about "reasonable overlap" is not particularly radical: the claim is simply that at some point (anomolous cases aside) what terms designate becomes *rigidified*.

The problem with the chemists' approach in the water-salt example is not that they have incorrectly analyzed the chemical composition of *S*. Instead, they have made a prespecialist categorization mistake by identifying the wrong substance in need of chemical analysis: they chose to analyze gross macrophysical substance *S*, which is rigidly designated as 'salt', rather than substance *W*, which is rigidly designated as 'water', on the incorrect presupposition that anything with the chemical composition of *S* is designated 'water'. And this is much like the proposed problem with Andreasen's account of race. The objection presented above is not that there is faulty biology, or even faulty analysis of that biology, but, rather, that there is a pre-specialist categorization mistake. The groups offered up as races (for which we can grant that scientists have identified real biological properties) simply are not races, just as salt is not water.

So the causal theory of reference will not bolster Andreasen's cladistic approach.[25] To put it roughly, the causal theory holds that natural kind terms get "baptized" in an initial naming process, and then those terms rigidly designate the objects that they so name. If, later, it is discovered that the baptizers (or other competent language users) identified properties with the object that are not, in fact, constitutive

[25] For her part, Andreasen holds that her approach "does not depend on" the causal theory—see "Race," p. S662.

of the object, so much the worse for the folk definitions. On this theory, science, rather than folk usage, tells us what properties are associated with the term in question. The reason this does not aid Andreasen's approach is that her theory of race requires de-rigidifying the term 'race'. On that approach, 'race' no longer picks out the same macrophysical objects (say, the three major races), nor does it pick out a reasonably overlapping object, which subsequently could be determined by scientists to have a different underlying structure than competent language users previously thought.[24] Rather, Andreasen's approach picks out different objects entirely (the nine populations in Figure 2).

In short, then, it does not help Andreasen's cladistic approach to privilege specialist over nonspecialist understandings of race. For, first, the specialist understanding of race (as presented by Andreasen) does not reasonably overlap with the folk category; that is, the relationship between cladistic 'races' and folk races seems closer to the relationship between salt and water than that between H_2O and XYZ.[25] Second, in any case, unlike water, there is no scientific consensus that race is a real biological kind; there is no expert consensus about (a) what entities we should identify as races (the three major races, Andreasen's nine, and so forth), or (b) whichever entities we choose, whether they have a biological basis.

V

All of the above does nothing to dispel the idea that if we could formulate a cladistic model that matches the common-sense notion of 'race', race would be biologically real. As it turns out, Kitcher presents a biological picture of race that ends up being strikingly

[24] Nor, for that matter, is it like Putnam's case of lemons that have changed from yellow to blue, since the proposed theory of race is not that the original objects picked out as races have themselves undergone a constitutional change, like lemons that have changed from yellow to blue—see Putnam, "Is Semantics Possible?" reprinted in his *Mind, Language, and Reality. Philosophical Papers*, Volume 2 (New York: Cambridge, 1975), pp. 139–52.

[25] This is worded too strongly: perhaps there is an argument for seeing the relationship between Andreasen's race and the folk concept as closer to that between H_2O and XYZ, rather than salt and water. But, as I argued above regarding the analogy with whales, this is not evident, and it requires a non-question begging argument, rather than mere stipulation. Andreasen does offer one analogy in response: like the difference between the folk concept of race and Andreasen's concept, a similar difference is found with species and higher taxa, insofar as they are defined historically by specialists, rather than according to shared phenotypic traits; and, since we would not therefore decide that species or higher taxa do not exist, so we should not decide that races do not exist—"Race," p. S665. The case of species, however, is different from the case of race: 'species' is not de-rigidified in the way that 'race' has been by Andreasen's account.

similar to everyday usage. For Kitcher, a pure race (where purity merely connotes reproductive isolation, rather than any kind of racial superiority) is just a subset of *Homo sapiens*, where offspring are of race *R* when their parents are of race *R* and parents are of race *R* when their offspring are of race *R*. Like Andreasen's account, this identifies races not "on the basis of traits," but, rather, on "patterns of descent." Finally, two further conditions are necessary for this classification to have any "biological significance." First, the members of the pure races must "have some distinctive phenotypic or genetic properties." Second, the mixed-race population cannot be so large that once-existing pure races are no longer reasonably substantial parts of the general population (*op. cit.*, pp. 92–94).

So far, this account is much like Andreasen's. Indeed, Kitcher identifies inbreeding among populations—reproductive isolation—as the factor that ensures the required difference between interracial and intraracial genetic properties. A key point of distinction from Andreasen's picture, however, is that for Kitcher, reproductive isolation persists through the present. On the basis of what he admits to be limited data, Kitcher finds that there are comparatively low rates of sexual union among blacks and whites, though the same is not true of whites and (at least some populations of) Asian Americans (*op. cit.*, pp. 99–100). On this account, human reproductive isolation is not a matter of geographic isolation (though it once was); rather, it is a function of what can be extremely subtle isolating mechanisms, such as cultural barriers to interracial relationships and breeding (*op. cit.*, pp. 105–10). As for the claim that there have been periods of high rates of interracial reproduction in the U.S. (in particular, the widespread rape of black women by white slaveowners), Kitcher notes those high rates were nevertheless much lower than intraracial breeding; and, importantly, we can conceive of this gene pool modification as a "coercive restructuring of the minority race" (*op. cit.*, p. 102). Indeed, since the mechanisms that isolate breeding populations are in this way socio-cultural, for Kitcher race is both socially constructed and biologically real.

This picture, then, is much more faithful to the folk category 'race' than Andreasen's: unlike Andreasen's model, Kitcher (a) finds reproductive isolation, and so race, persisting through the present; and (b) identifies as the races (isolated breeding groups) traditional racial groupings, such as black, white, and so forth.

Nevertheless, I do not think that this picture is sufficiently faithful to the folk category. Kitcher acknowledges that he has only accounted for a difference between Africans and Caucasians in the U.S., while he has not provided evidence for a distinction between Asians and

Africans and little for a division between Asians and Caucasians.[26] Moreover, since this notion of race hinges only on reproductive isolation, different social classes that are reproductively isolated (such as landowners and peasants in England after the Norman conquest) would end up being classified as different races (*op. cit.*, p. 103).

Accordingly, this presents a classification of 'race' that is substantially different from the folk meaning of the term. Kitcher's treatment of the folk category 'Asian' is particularly inconsistent with folk usage. If there is no division between Asians and Caucasians or Asians and Africans, but there is a division between Africans and Caucasians, what happens to Asians in locations where the three meet (*op. cit.*, p. 100)?[27] Are Asians raceless? Or perhaps Asians become Caucasian, or African? Or do Asians combine with either Africans or Caucasians to form a new, fourth race? It is difficult to see how any of these options would be sustained, or how dramatic political implications could be mitigated. More to the point, however, none of them seems to match the folk categorization of race.

One might argue, however, that this problem arises only because of insufficient data on reproductive patterns. That is, if we had more complete data, we might find significantly higher rates of intraracial reproduction than interracial reproduction; and, since these reproductive behaviors are influenced by people's *perceptions* of what race is (rather than any biological facts), the resulting breeding populations are going to end up matching folk racial groupings.

This question leaves us waiting for more complete data. In the meantime, however, Kitcher's account faces two further problems. First, Kitcher himself admits that to be "a workable biological conception of race," there must be mating patterns between the races that are sufficient "to sustain the distinctive traits that mark the races (which must, presumably, lie, at least in part, in terms of phenotypes, since organisms have no direct access to one another's genes)" (*op. cit.*, p. 97). That is, while Kitcher's model identifies races as reproductively isolated breeding populations, and so phenotype alone is not the *basis* for race, this reproductive isolation is *significant* only so long as it maintains phenotypic differences between the races. But this exposes his model to old worries about the possibility of making sense of racial

[26] Kitcher does not mention Native Americans, and only briefly mentions Hispanics, in remarking that there are low rates of sexual union between Hispanics and Asians.

[27] For Kitcher, since interracial sexual unions may occur at different frequencies at different locations due to different isolating mechanisms at those locations (say, differences between Oakland and Memphis in the cultural barriers to interracial dating), racial distinctions may shift depending on the locale

divisions based on phenotype. Naomi Zack puts it this way: "The visual and cultural markers for membership in the black race differ too greatly for there to be any physical traits shared by all black individuals, and likewise for whites."[28] By Kitcher's own standards, then, in order for race to be a significant category, we at least need an additional story about which phenotypic traits are supposed to go with which races. If Zack is right, such a story cannot be told.

Second, the races-as-breeding-populations model seems too broad, as evidenced by the counterintuitive result that peasants and landowners would have to be considered races. Folk usage, of course, distinguishes between socioeconomic classes (for example, peasants) and races (for example, "white people"). And this reflects an intensional difference in the two meanings of race: the folk concept of race seems to include a phenotypical component, including traits such as skin color, that does not correlate one-to-one with class status; class status is orthogonal to membership in any given race.

The more general point here is that potentially there are many breeding populations (based not only on class, but small regions, professions, cultures, and so forth), which are not accurately labeled 'races'. Indeed, it is possible that nonracial breeding populations even could generate distinct phenotypic features, such that one population has, say, "hitchhiker's thumb." On Kitcher's model, such a population would have to be called a 'race', which seems to stretch the meaning of 'race'. Thus, while Kitcher's account of race might seem to avoid significant divergence from the folk category of race, in the end there do seem to be substantial intensional and extensional differences between the two.

VI

Genetic findings recently published in *Science* by Noah Rosenberg and others might seem capable of plugging the holes that we have so far seen in the races-as-breeding-populations theory.[29] They report that while 93–95% of genetic variation occurs within geographic populations, a further 3–5% of genetic variation distinguishes five populations that correspond to five "major geographic regions": Africa, Eurasia, East Asia, Oceania, and America (where Eurasia includes Europe, the Mid-East, and Central/South Asia). In essence, without prior identification

[28] Zack, "Life after Race," in Zack, ed., *American Mixed Race: The Culture of Microdiversity* (Lanham, MD: Rowman and Littlefield, 1995), pp. 297–307, here p. 303.

[29] Rosenberg, Jonathan K. Pritchard, James L. Weber, Howard M. Cann, Kenneth K. Kidd, Lev A. Zhivotovsky, and Marcus W. Feldman, "Genetic Structure of Human Populations," *Science*, CCXCVIII (December 20, 2002): 2381–85.

of one's geographic ancestry, genetic information can be used to identify that ancestry. Given that some disease risks are higher for different populations, this is particularly significant for health care treatment and epidemiological research.

It is notable that the authors of this study never use the word 'race' to describe the geographical populations they identify. Yet *The New York Times* reports that they were willing to say in interviews that "[the five major] regions broadly correspond with popular notions of race."[30] Can we therefore say that this is the biological basis for race?

While these data certainly get us closer to the folk concept of 'race', it still seems too distant to say that race is biologically real. Consider, again, both the intension and the extension of the folk concept. Throughout, I have emphasized the significant degree to which the intension of 'race' is tied to gross morphological features, such as skin color. The data reported by Rosenberg and his co-authors provide no indication that those features can be mapped onto population-based properties, and, again, there is the point from Zack that even if such data becomes available, it is difficult to see how we would identify even vague criteria for assigning certain phenotypic properties to one race and not to another.

Relatedly, the geographic populations identified by Rosenberg and his co-authors seem extensionally different from the folk notion of race. For example, consider that the population of Adygei, from the Caucasus (from which the term 'caucasian' originates), is lumped into the same major geographical population as the French, Palestinians, and Pathan/Pushtuns of Afghanistan. Perhaps we should say that these groups all compose one race, but I think that further argument is required for doing so. This point becomes particularly compelling when one considers the political implications of these categories. Might, for instance, a race-conscious Palestinian categorize herself within a group that does not include the French?

Political questions are important here not just because possible answers reveal something about the way we categorize ourselves. In section III, I noted that the new racial biological realism needs a further argument showing either that we should replace race with the distinct (and more biologically defensible) concept of breeding

[30] "Gene Study Identifies 5 Main Human Populations," *The New York Times* (December 20, 2002), Late Edition, Section A, page 37. It should be noted that geographic populations are not identical to breeding populations. In defense of the breeding population model, however, one might argue that reproductive isolation results from geographic barriers. A claim of this sort is made in the *Times*' report on Rosenberg et al., "Genetic Structure."

population, or that we should modify our conception of race in this substantial way. Such a position bears the burden of explaining how it would be practically possible to revise so significantly entrenched racial discourse.[31]

But in addition to the practical problems that would arise, there are also significant political hurdles facing any argument for either replacing race with another discourse entirely or substantially modifying the concept of race. For example, the classifications provided by both Kitcher and Andreasen struggle to make sense of the folk category 'Asian'. If our political practices ought to contain no descriptive falsehoods, then their accounts of race would disallow political tools focused on Asians, such as those that might fight uniquely anti-Asian discrimination. This is only one of the problems with the "black/white binary" model of race, which others have discussed more extensively.[32] Furthermore, if the races-as-breeding-populations account ends up conflating different political axes, such as race and class (for example, English peasants), we might be left with impoverished political resources for dealing with social problems that are unique to each (which, again, others have discussed more extensively). The point can be stated briefly: a substantive revision to, or replacement of, the concept 'race' must show either that it can offer the requisite conceptual resources to justified political causes, or that those causes must be abandoned for the sake of conceptual coherence.

Yet one might take a different argumentative tack and argue that this substantive revision is already taking place, rendering any political or practical questions moot. After all, since scientific discourse is not wholly isolated from folk discourse, it is conceivable that the biologists' identification of breeding populations is itself changing the meaning of the folk concept of race, particularly in light of significant media coverage of these new biological data. While this change is conceivable, however, it seems doubtful that it has already been effected, given the divergence between the folk concept and the concept of races as breeding populations that has been discussed here. The source of this divergence is evident: the folk meanings of socially

[31] One potential practical problem with the revisionist program is that such a revised concept may retain some hidden references to the previous, inadequate concept. Appiah expresses concern about this in "Social Forces, 'Natural' Kinds," in Abebe Zegeye, Leonard Harris, and Julia Maxted, eds., *Exploitation and Exclusion: Race and Class in Contemporary US Society* (London: Hans Zell, 1991), pp. 1–13, in favor of abandoning race-talk. "if you want to talk about morphology, talk about morphology; if you want to talk about populations, talk about populations" (p. 12, n. 9)

[32] See Linda Martín Alcoff, "Latino/as, Asian Americans, and the Black-White Binary," *The Journal of Ethics*, VII (2003): 5–27.

charged concepts like race are influenced by social practices (which, in the case of race, have significant normative dimensions) that often are not themselves restricted to being biologically accurate, as much as we might like them to be. Thus, given the divergence between the folk concept of 'race' and the biological facts, racial discourse seems to be at a crossroads: either we must acknowledge that we have no biological basis for that discourse, or the meaning of 'race' must be changed to reflect the biology. In this case, the pressing question is one of social policy: Should such a change in racial discourse be effected? The suggestion here is that any argument in favor of this substantive revision requires a substantive defense against the potential political problems just raised.

VII

All of the above leaves open the possibility that there might be some as yet undiscovered biological basis for race (though any such account of race would need to respond to Zack's challenge that there simply are no phenotypic traits shared by all blacks or all whites). It also leaves intact the idea that we might profitably make distinctions between humans based on reproductively isolated breeding groups, even if we do not cash out race in these terms. Finally, the arguments made here leave open the possibility that race might be real as a social, if not biological, kind. The foregoing, however, does suggest that we have not yet been given an adequate argument for holding that breeding populations are the biological basis of race.

JOSHUA M. GLASGOW

California State University/Bakersfield

COMMENTS AND CRITICISM

THOMSON AND THE SEMANTIC ARGUMENT AGAINST CONSEQUENTIALISM*

In "The Right and the Good," Judith Jarvis Thomson[1] argues that important normative conclusions can be derived from the correct account of the meaning of 'good'. She argues in particular against consequentialism on the basis of the semantic thesis that "all goodness is goodness-in-a-way."

In this essay, I will argue that Thomson's attack on consequentialism is less powerful and less precisely targeted than she supposes. For we can develop an argument against pure obligation or categorical imperatives that is largely parallel to Thomson's argument against pure goodness. The most plausible response to this largely parallel argument is to allow that the existence of pure obligation is neither semantically ruled out nor semantically guaranteed. Similarly, the most plausible response to Thomson's argument against pure goodness is that the existence of pure goodness is neither semantically ruled out nor semantically guaranteed. I will argue, however, there is something importantly right about Thomson's attack: she does succeed in undermining one important source of support for consequentialism.

I. THE ATTACK ON CONSEQUENTIALISM

Thomson's attack on consequentialism is in the tradition of an attack earlier developed by Philippa Foot.[2] Thomson's argument can help-

* I am grateful to Bill Nelson, Alastair Norcross, and audiences at UTIL 2000 and the University of Houston for most helpful comments on earlier versions, and to Judith Thomson for generously supplying me with copies of the proofs of *Goodness and Advice* prior to publication.

[1] This JOURNAL, XCIV, 6 (June 1997): 273–98 (hereafter RG) My primary focus will be this article, but Thomson's views on these issues have also been developed in a number of other places, and sometimes it will be important to draw on these other presentations Other related pieces are: "On Some Ways in Which a Thing Can Be Good," *Social Philosophy and Policy*, IX, 2 (Summer 1992): 96–117, "Goodness and Utilitarianism," *Proceedings and Addresses of the American Philosophical Association*, LXVII, 2 (October 1993). 145–59—also LXVII, 4 (January 1994): 7–21 (page references to the latter); Gilbert Harman and Thomson, *Moral Relativism and Moral Objectivity* (Cambridge: Blackwell, 1996), pp. 125–54 (which I will refer to as "Relativism and Objectivity"); *Goodness and Advice* (Princeton: University Press, 2001).

[2] See Foot's "Utilitarianism and the Virtues," *Mind*, XCIV (1985): 196–209 (also presented as an APA presidential address). Thomson discusses the source of the argument in Foot in RG 275, note 3. Foot gives to my mind a less thorough and plausible account of the semantics of goodness than does Thomson. Samuel Scheffler

0022-362X/03/0009/475–86 © 2003 The Journal of Philosophy, Inc.

fully be seen as involving two key steps. The first step is a semantic claim about goodness, the claim that "all goodness is goodness-in-a-way." The second step is the application of this claim about the nature of goodness to undermine consequentialism.

Begin with the semantic claim:[3]

> All goodness is goodness in a way...if we do not know in what way a man means that a thing is good when he says of it, 'That's good', then we simply do not know what he is saying of it (RG 276).

Thomson's version of this claim is significantly different from some earlier versions defended by Peter Geach, Paul Ziff, and others.[4] Geach took it that 'good' always applies to a kind (for example, 'good hammer', 'good person'), and that to know the standards involved one needed to know the kind. Thomson claims instead that, "The role of 'good' as a modifier on a kind-term is parasitic on the role of 'good' modified."[5] 'Good' occurs "modified" in such expressions as 'good at dancing' or 'good to use to hang things'.

Nonetheless, for Thomson, the arguments for the claim that all goodness is goodness in a way are arguments that go back to Geach and Ziff. Thomson does not make these arguments particularly explicit, taking it that these earlier thinkers have established the central idea. As an attempt to make the argument for this central idea more explicit, try this: 'good' occurs in many different contexts, applied to many different kinds of things, things good for quite different reasons.

---

responds to Foot in "Agent-Centered Restrictions, Rationality and the Virtues," *Mind*, XCIV (1985): 409–19. As he characterizes his response in the introduction to his collection, *Consequentialism and Its Critics* (New York: Oxford, 1988), p. 12, Scheffler argues that "the seeming irrationality of agent-relative morality does not depend on the assumption that it always makes sense to talk about better and worse states of affairs. The conflict goes deeper than that...it is not so easy to dispel the appearance of a conflict between agent-relative constraints, on the one hand, and a very powerful and familiar conception of rationality on the other." I sympathize with his reconstruction of the paradoxical character of agent-relative morality, but do not think, for reasons suggested in section IV, below, that the possibility of giving such a reconstruction is itself sufficient to vindicate consequentialism.

[3] Although Thomson presents sometimes semantic and sometimes metaphysical versions, surely the claim should properly be understood as initially semantic. The metaphysical claim should be understood as inferred from the semantic claim. Otherwise Thomson will just be begging the question. One might suppose that "all goodness is goodness-in-a-way" is a formulation of the inferred metaphysical claim rather than of the semantic claim, but I take Thomson to employ it as a formulation of the semantic claim—see especially, "Goodness and Utilitarianism," p. 11, where she says that "all goodness is goodness-in-a way" is a more "compact" way of putting the claim that 'good' is incomplete.

[4] See Geach, "Good and Evil," *Analysis*, XVII, 2 (December 1956): 33–42; Ziff, *Semantic Analysis* (Ithaca, NY: Cornell, 1960), chapter 6.

[5] "Goodness and Utilitarianism," p. 9.

So surely it is overwhelmingly plausible that 'good' shares with standard indexicals the feature Thomson calls *incompleteness*: when someone asserts something of the form '*x* is good', unless she or the context tells us what way of being good or standards she has in mind, we will not know what she is saying about *x*.

Now consider the application of the semantic claim to undermine consequentialism. The consequentialist accepts what Thomson calls, after G.E. Moore, "Moore's story" about the right and the good.

> There is such a property as goodness...[and].. there is also such a relation as being better than, or *betterness*...Moore's story then concludes: the right is analyzable in terms of the relation betterness. For it to be the case that Alfred ought to do a thing at a time is for it to be the case that the world will be better if he does the thing than it will be if he does any of the other things it is open to him to do at the time (RG 273–74).

But if the semantic claim is correct, Moore's story is mistaken.

> If there is no such property as goodness, and no such relation as betterness, then, a fortiori, the right is not analyzable in terms of the relation betterness.... What is wrong with Moore's story is...its very beginning (RG 276).

In earlier presentations, Thomson goes on to consider and reject possible utilitarian responses, responses that involve specifying the way of being good the consequentialist has in mind. We will consider these issues in section IV. What we have already presented, though, is the heart of Thomson's initial argument.

II. A PARALLEL ARGUMENT AGAINST PURE OBLIGATION

The thesis that all goodness is goodness-in-a-way presents problems for consequentialism. So far, we have an argument against consequentialism. But, I will argue, a largely parallel linguistic thesis about 'ought' is equally plausible. The reasons we have to think that all goodness is goodness-in-a-way are also reasons to think that all 'oughts' are 'oughts'-of-a-kind. Establishing these theses renders problematic a crucial deontological notion: the notion of a categorically binding 'ought' or categorical imperative. So Thomson's argument against consequentialism threatens a concept central to one of its main rivals.[6]

As we did in presenting Thomson's argument, it is helpful to distin-

[6] I do not mean to imply here that *only* deontologists make use of the notion of a categorical 'ought'. Certainly they are among those who do; to the extent that others do as well, my overall point—that Thomson's strategy of argument has radical implications extending well beyond consequentialism—is reinforced.

guish two steps in the parallel argument. The first step is to establish a semantic claim about obligation and 'ought', the claim that all 'oughts' are 'oughts'-of-a-kind. The second step is to employ this semantic claim to raise problems for the idea of a categorical imperative.

Begin with the first step. The evidence for the incompleteness of utterances involving 'ought' is just as compelling as the evidence for the incompleteness of utterances involving 'good'. Consider the sentence 'The car ought to be fixed'. Without knowing more about the context, we do not know whether this sentence expresses the expectation that the car will by now have been fixed, the prudential judgment that it would be best for the owner to have the car fixed, the moral judgment that someone else owes it to the owner to fix the car, or the judgment that the car is in decent enough shape that it is worth someone's fixing. Further contextual information is needed to decide between these (and probably other) possibilities. To paraphrase what we said earlier about 'good': 'ought' occurs in many different contexts, applied to many different kinds of things, expressing many different kinds of obligation. So surely it is overwhelmingly plausible that 'ought' too is incomplete: when someone asserts something of the form '*x* ought to be done', or '*A* ought to do *x*', unless she or the context tells us what kind of obligation she has in mind, we will not know what she is saying. Moreover, just as Thomson can appeal to a significant philosophical tradition[7] of holding that 'good' is incomplete, so there is also a significant philosophical tradition[8] according to which 'ought' is incomplete. If all goodness is goodness-in-a-way, all 'oughts' are 'oughts'-of-a-kind.

It is not necessary to establish that 'ought' works in *precisely* the same way in which 'good' works. It does not. Among the important differences are these: 'ought' is a verb; 'ought' (more than 'good') suggests the unique correctness of some option; unlike 'good', 'ought' does not *always* pick out systems of norms or standards; sometimes, instead, it picks out considerations that are propositional, laws or facts, not standards.[9] (To see this, consider the difference between

[7] Including Geach and Ziff, Foot herself, and G. H. von Wright in *The Varieties of Goodness* (New York: Humanities, 1963)

[8] Including: Roger Wertheimer in *The Significance of Sense* (Ithaca, NY: Cornell, 1972); Harman in various writings beginning with "Moral Relativism Defended," *The Philosophical Review*, LXXXIV (1975): 3–22; David Wong in *Moral Relativity* (Berkeley: California UP, 1984), and, most recently, Peter Unger in *Living High and Letting Die* (New York: Oxford, 1996), chapter 7.

[9] On this point, see Wertheimer, pp. 89–92. He distinguishes "systems of ideality" from "systems of actuality"; the difference is (roughly) a matter of direction of fit.

'oughts' of expectation (for example, 'she ought to be here by now' and deontic 'oughts' (such as, 'you ought not to tell lies'). The former do not seem to invoke norms, whereas the latter do.)

But no such precise correspondence between 'good' and 'ought' is required for the argument. All that is required is that 'ought', like 'good', is incomplete. And, whatever the further semantic complications, the arguments for this incompleteness claim are surely overwhelming.

We now move to the second step in the argument—the application of the idea that all 'oughts' are 'oughts'-of-a-kind to undermine deontology. Thomson claims that if all goodness is goodness-in-a-way, then there is no such thing as pure unadulterated goodness. But then, if all 'oughts' are 'oughts'-of-a-kind, there is no such thing as pure unadulterated obligation, no such thing as what one just plain ought or ought period to do. That is, there is no such thing as a categorical imperative.

Let us take this a little more slowly. The central idea in the second stage of this argument is the idea that a plain 'ought', a categorical 'ought', an 'ought'-period is *not* just an 'ought' of a certain kind. So if all 'oughts' are only 'oughts'-of-a-kind, then semantics leaves us no room for a categorical ought.

To help explain the contrast between categorical and non-categorical 'oughts', consider the following passage from Stephen Darwall's[10] philosophical remembrance of William Frankena:

> In "On Saying the Ethical Thing" (1966), Frankena ventured the thought ..that normative judgments are always made relative to some assumed perspective or point of view, and that the rationality that normative judgments claim for the attitudes they standardly express is always within that point of view.
>
> [However]...'oughts' that seem binding to us from the moral point of view, present themselves (from that point of view) as binding nonrelatively, not just binding relative to that perspective...moral obligations present themselves as categorically binding (*ibid.*, pp. 698–700).

Thus there is a contrast between perspectival reasons and categorically binding genuine reasons. The argument we are here developing, parallel to Thomson's argument considered in section I, claims that it follows from the thesis that all 'oughts' are 'oughts'-of-a-kind that

[10] "Learning from Frankena: A Philosophical Remembrance," *Ethics*, CVII (July 1997): 685–705.

there are no categorically binding reasons. All there are are perspectival reasons.

Any such conclusion will have important implications for moral theory. At the very least, the views of Kant and broadly Kantian theorists (like Darwall) will be undermined by it. They take the notion of a genuine, nonperspectival reason to be central to moral thinking and indeed to all rational deliberation. Moreover, and crucially for the critique of Thomson, the parallel argument will undermine Thomson's own view about 'ought' developed in a number of places, most prominently in the second lecture of *Goodness and Advice*. Thomson there suggests that there is only one, categorical, advice sense of 'ought' (*op. cit.*, pp. 43–82).[11]

But the argument to the claim that there is no pure obligation from the thesis that all 'oughts' are 'oughts'-of-a-kind is just as plausible as the argument to the claim that there is no plain goodness from the thesis that all goodness is goodness-in-a-way.[12] So Thomson's attack on consequentialism, like Bernard Williams's before it,[13] threatens some of consequentialism's main rivals just as much as it threatens consequentialism, and indeed threatens Thomson's own positive view.

III. RESPONSES TO THE PARALLEL ARGUMENT AGAINST PURE OBLIGATION

At this point, we should be suspicious. How can what seems to be a semantic thesis really undermine a concept central to moral theory (as at least some philosophers conceive it)? I will argue that our suspicions are justified, and that the thesis that all 'oughts' are 'oughts'-of-a-kind does not undermine the idea of categorically binding reasons. And once we have seen this, I will argue, we will also see that the thesis that all goodness is goodness-in-a-way does not undermine the possibility of understanding the kind of goodness the consequentialist needs.

Before considering what I take to be the right response, it is impor-

[11] Thomson discusses the distinction between categorical and hypothetical imperatives on pp. 48–49.

[12] It is striking that there is a point at which Foot is tempted in effect by the parallel argument against pure obligation. In a footnote on p. 169 of "Morality as a System of Hypothetical Imperatives" in *Virtues and Vices* (Berkeley: California UP, 1978), Foot writes: "This 'ought'—the one in the sentence 'One ought to be moral'—is supposed to be free floating and unsubscripted, and I have never found anyone who could explain the use of the word in such a context."

[13] For the "integrity objection" narrowly targeted against utilitarianism, see Williams's contribution to J.J.C. Smart and Williams, *Utilitarianism: For and Against* (New York: Cambridge, 1973); for the objection targeted much more broadly, see *Ethics and the Limits of Philosophy* (Cambridge: Harvard, 1985)

tant to consider Thomson's own response. Though she does not articulate the argument considered in section II above, she does explicitly deny the semantic parallel between 'good' and 'ought'. She introduces two versions of the idea that there is such a parallel—the idea that 'ought' is three ways ambiguous, between 'oughts' of expectation, hypothetical 'oughts', and categorical 'oughts', and the idea that 'ought' is multiply ambiguous. She rejects both proposals. Since the latter is the proposal that provides the first step in the argument developed in section II, I will focus just on her arguments against it.

In *Goodness and Advice*, Thomson gives two reasons for thinking that "the multiple ambiguity idea" is "a very bad idea." First "it is wildly implausible to think that the word 'ought' is multiply ambiguous in that way":

> Suppose that Alfred's paying Bertha five dollars would be good for Bertha but bad for Alfred. Alfred asks whether he ought to pay Bertha five dollars. A proponent of the Multiple Ambiguity Idea says that the only advice we are in a position to give him is that in the 'ought$_{\text{goodness for Bertha}}$' sense of "ought" Alfred ought to pay Bertha, whereas in the 'ought$_{\text{goodness for Alfred}}$' sense of "ought," it is not the case that Alfred ought to pay Bertha. Surely there remains a further question, namely whether Alfred just plain ought to pay Bertha. In asking the question whether Alfred...ought to pay Bertha, we are using the word "ought" in *the* advice sense (*op. cit.*, p. 46).

The trouble with this argument is that it is hard to see why Thomson's insistence that Alfred is asking a real question carries any more weight than the parallel insistence of Moore. As Thomson characterizes him earlier in *Goodness and Advice*,

> Moore said it is clear that some things are good, some are bad, and some are neither. Goodness, he said, is the property all and the only good things have in common. That is the property that we would be ascribing to a thing—whether an event or anything else—if we said of it, "That's good"; and that is the property such that we are asking whether a thing possesses it when we ask about the thing "Is it good?" (*op. cit.*, p. 17).

Proponents of Thomson's view (which one might call the *multiple ambiguity idea*) about goodness will deny evidential weight to Moore's insistence that there is a genuine question whether a thing is just plain good. Proponents of the multiple ambiguity idea about 'ought' will deny evidential weight to Thomson's insistence that there is a genuine question whether Alfred just plain ought to pay Bertha. No reason I can see has been supplied to credit Thomson but disbelieve Moore.

The second reason Thomson gives to reject the multiple ambiguity idea is that "it gives moral goodness and badness the wrong role in assessments of what a person ought to do" (*op. cit.*, p. 46). Immediately after alluding to this second reason, Thomson says "I will return to moral goodness and badness later" (*op. cit.*, p. 46). Presumably she here refers to her later account of the role of injustice and miserliness in determining the truth value of 'ought' judgments. She says that "to act unjustly is to fail to do what another person has a *non-overridden* right to one's doing" (*op. cit.*, p. 59). She then develops the concept of *miserliness*: "I will describe as 'miserly' any act that provides its agent with a small gain...or even no gain at all, despite the act's being very bad for another person" (*op. cit.*, p. 61–62). Both are, she claims, second order ways of being bad, because the answers to the questions whether actions are unjust or miserly turn in part on the answers to questions about first order goodness and badness, specifically questions about how good and bad those actions are for people. Thomson then includes among the principles that determine the truth value of 'ought' judgments:

(III) *A* ought to *V* if his not *V*-ing would be either unjust or miserly, and *A* ought not *V* if his *V*-ing would be either unjust or miserly (*op. cit.*, p. 67).

Thomson's second reason to reject the multiple ambiguity idea presumably then amounts to the claim that (III) is correct, rather than its being true that, as the multiple ambiguity idea would suggest, *A* ought-morally to *V* if and only if his *V*-ing would be morally good. Again, it is hard to see this reason as having any force. To accept Thomson's (III) is to take a particular stand on a quite controversial issue about the relation between morality and practical reason. It is to hold not only that there are valid categorical 'oughts', but also that in determining what we ought to do moral considerations override, in the way (III) specifies, other considerations. Either sub-claim could be and has been denied. Some disbelieve in any such valid categorical imperatives; others allow that there are valid such imperatives but deny that moral considerations are rationally overriding. Thomson's is certainly a possible and defensible view about these issues, and certainly conflicts with the multiple ambiguity view. But it is hard to see how, in sketching it as she has, she has given any strong reason to believe it or to reject the alternative multiple ambiguity view.

So Thomson fails to show that we should reject the multiple ambiguity idea, the first premise of the parallel argument. The way to respond to the argument that there are no categorically binding reasons because all 'oughts' are 'oughts'-of-a-kind is not to deny the semantic

premise. It is instead to show that there is a kind of ought that *is* the kind that generates or constitutes categorically binding reasons. Two obvious suggestions present themselves: categorically binding 'oughts' are *moral* 'oughts', or they are *rational* 'oughts'.

Consider first the idea that categorically binding 'oughts' just are moral 'oughts'. Interestingly, this is an idea that Thomson in effect endorses in her brief discussion of categorical imperatives in *Relativism and Objectivity* (*op. cit.*, pp. 148–50). As she understands it, to say that

(O): '*A* ought to do *X*'

is a categorical imperative is just to interpret (O) as saying

(OC): 'Morality requires *A* to do *X*'

To assess this idea, that categorically binding 'oughts' are moral 'oughts', we need to focus on the term 'morality' in OC. How are moral requirements to be identified? One suggestion would be that moral requirements are just requirements with a distinctive content and role—for example, impartial requirements designed to overcome social conflicts. But then notice that if we understand the term 'morality' in (OC) this way, we do not have a categorically binding ought. We do not have an imperative necessarily applying to and binding on everybody. How else might moral requirements be identified? No identification will be satisfactory if it fails to include this special kind of necessity and bindingness. But this special sort of necessity and bindingness is just what is involved in the notion of an 'ought' of reason.

This suggests that categorically binding 'oughts' are better understood in the second way as 'oughts' of reason or rationality. Darwall suggests a way of interpreting such 'oughts': roughly, a categorically binding reason is a consideration a person would be moved to act on if she reflected rationally about it.[14] To avoid circularity, the ideal of rational reflection must then be an independent, procedural or formal, not a substantive one.

What Darwall suggests here is controversial. One might wonder how possible it is to specify in a neutral and plausible way the procedural or formal ideal of rational reflection or deliberation. And if this ideal can be specified reasonably adequately, one might wonder how determinate the categorically binding reasons that emerge will be. But

[14] Darwall calls this view "constitutive existence internalism." For a classification of types of internalism, see his "Internalism and Agency," *Philosophical Perspectives*, VI (1992): 155–74. Compare also Christine Korsgaard, "Skepticism about Practical Reason," this JOURNAL, LXXXIII, 1 (January 1986): 5–25

surely Darwall does enough to show what kind of 'oughts' categorically binding 'oughts' will be, if there are any. He does so by spelling out the theoretical role of and constraints on such categorically binding 'oughts'.

Thus the argument that there are no categorically binding 'oughts' because all 'oughts' are 'oughts'-of-a-kind fails. It fails because we can explain what such categorically binding 'oughts' would be like if there were any. To put this another way, it fails because we can explain how categorically binding 'oughts' are construed by the theories according to which there are such categorically binding 'oughts'. To say this is not to say that such theories are correct and that there are such categorically binding 'oughts'. It is only to say that such theories *may* be correct; whether they are and how that could be shown are questions which we cannot adequately address here. Their correctness cannot be semantically guaranteed, but it also cannot be semantically ruled out.

IV. RESPONSES TO THOMSON'S ARGUMENT AGAINST CONSEQUENTIALISM

Once we see how the parallel argument fails, we will see that Thomson's argument against plain goodness fails in the same way. It *is* possible to explain what way of being good is being picked out by consequentialism, just as it is possible to spell out the kind of 'ought' involved in categorically binding reasons. Consequentialism too cannot be semantically ruled out and cannot be semantically guaranteed.

Consider first the possibility that consequentialism might be spelled out in terms of *intrinsic* goodness, as Michael Zimmerman[15] suggests. As he spells the suggestion out:

> The key question then is: *what* way of being good does being intrinsically good involve?... The answer is: *ethical* goodness. When it is said that pleasure, or knowledge, or beauty, or virtue is *intrinsically* good...what is meant is that all these things are *ethically* good.... That which is intrinsically good is preferable to that which is not, the '-able' here expressing *moral* worthiness (*ibid.*, pp. 404–05).

Nothing in Thomson's argument rules out the possibility of understanding goodness this way. That there is such a thing as intrinsic

[15] See "In Defense of the Concept of Intrinsic Value," *Canadian Journal of Philosophy*, XXIX, 3 (September 1999): 389–410. I find much of Zimmerman's critique plausible. His approach is quite different from the one I take here: he does not exploit the parallel between 'good' and 'ought'; he focuses on vindicating the concept of intrinsic value more than on Thomson's critique of consequentialism; and he does not distinguish the two possible aims of Thomson's critique as I do later in this section.

goodness of this kind cannot be semantically ruled out, though it cannot be semantically guaranteed.

Notice also a related possibility that Zimmerman does not exploit. By making the '-able' in "preferable" express not moral but rational worthiness, we would have specified a kind of goodness closer to the kind of obligation in terms of which, we argued above, categorical imperatives are best understood. And the reasons we considered above suggest again that this is the better way to understand intrinsic goodness. Thus we must reject Thomson's conclusion in "The Right and the Good," that "there is no such property as goodness."

This does not, however, mean that there is nothing to Thomson's critique. In earlier presentations, she sometimes has a more modest goal: not showing that consequentialism is incoherent, but instead undermining one important argument for it. In "Goodness and Utilitarianism," she notices that a utilitarian might accept that all goodness is goodness-in-a-way, but still accept a version of the utilitarian formula that one ought to do whatever is better, by picking some way of being good and claiming that one ought to do anything that is better in *that* way. She canvasses several possible ways of being good to insert into the utilitarian formula. The most plausible of these is goodness for sentient beings. Her treatment of this possibility is instructive:

> It is of great interest that something is lost to the utilitarian who makes this choice. As Philippa Foot said, the chief attraction of utilitarianism for many people is that it sounds like a necessary truth. Suppose it is clear that if I do such and such then the world that will come about will be better—just pure, unadulterated better—than the world that will come about if I do so and so. Surely, then, it is clear that I ought to do the such and such. How *could* morality direct me otherwise?... Well, if all goodness is goodness-in-a-way, then there is no such thing as one world's being pure, unadulterated better than another. Consider, then, the utilitarian...who fixes on goodness-for.... What emerges is that the theory we are supplied with by the utilitarian I envisage here is not at all plausibly viewable as a necessary truth. People may find it attractive, but if they do, it is a moral judgment which they are finding attractive, a—not at all obvious— moral judgment to the effect that it is goodness-for *alone* that matters to the question what a person ought to do (*op. cit.*, p. 13).

There is something right about her diagnosis of one source of consequentialism's appeal: the thought that the consequentialist principle cannot be rejected because it has the "flavor of necessary truth." And Thomson is right to reject the idea that consequentialism is a necessary truth. Just as the truth of consequentialism cannot be semantically ruled out, it cannot be linguistically guaranteed. Some

care is needed, however, in explaining why, though consequentialism is not a necessary truth, it might seem like one.

Let us formulate the consequentialist principle thus: the right action is the one with the best consequences. Accepting consequentialism involves accepting (i) that there is such a thing as the most valuable option, and (ii) that the proper response to value is to promote it rather than to honor it.[16] Some opponents of consequentialism reject (i), some reject (ii). However plausible (i) and (ii) may be, neither is a necessary truth. So consequentialism, which requires both (i) and (ii), is not a necessary truth.

The consequentialist principle will seem like a necessary truth, I suggest, only if we hear its assertion as pragmatically implying the truth of (i) and (ii). It is easy to see how this would happen with (i): we hear the assertion of the principle as guaranteeing what it presupposes, that there is such a thing as the best or most valuable option ((i)). It is harder to see why it would happen with (ii). But, I suggest, if consequentialism sounds like a necessary truth this must be because the assertion of the consequentialist principle is itself taken also to guarantee the correctness of an instrumental approach to value ((ii)).

V. CONCLUSIONS

Thomson's semantic argument against consequentialism from the claim that all goodness is goodness-in-a-way has broader implications than she recognizes. For a largely parallel semantic argument from the claim that all 'oughts' are 'oughts'-of-a-kind threatens to render incoherent the crucial deontological concept of a categorical 'ought'. Neither argument succeeds in showing that the concept it targets is incoherent. For it is possible to specify the way of being good the consequentialist has in mind and the kind of 'ought' the deontologist has in mind. Whether there really is such a thing as this way of being good or this kind of obligation cannot be shown semantically; establishing either will require vindicating the normative theory in which it is embedded. Thomson is, however, right when, more modestly, her aim is simply to reject one familiar kind of argument for consequentialism, an argument deriving from the apparent necessity of the consequentialist principle.

DAVID PHILLIPS

University of Houston

[16] Here I draw on Philip Pettit's discussion in the article "Consequentialism," in Peter Singer, ed., *A Companion to Ethics* (Cambridge: Blackwell, 1991), pp. 230–40.

BOOK REVIEWS

Kant's Theory of Taste: A Reading of the Critique of Aesthetic Judgment. HENRY E. ALLISON. New York: Cambridge University Press, 2001. 424 p. Cloth $75.00, paper $26.00.

Kant's Theory of Taste completes Henry Allison's trilogy on Kant's critical system.[1] Allison now offers a systematic account of the first division of Kant's *Critique of the Power of Judgment*: *The Critique of Aesthetic Judgment*.

The book is in four parts. The first part is an analysis of the two Introductions to Kant's *Critique of the Power of Judgment*. In chapter one, Allison gives a compelling explanation of the relation between the "merely reflective" aesthetic judgments of the third *Critique* and the broader notions of reflection and reflective judgment at work in the first *Critique*, namely in Kant's theory of cognition. In chapter two, he offers an interpretation and defense of Kant's claim that one and the same "principle of the purposiveness of nature for judgment" underlies our empirical judgments about objects of nature, on the one hand; and the "merely reflective" judgments of taste, on the other hand.

The second part of the book is a detailed commentary on the Analytic of the Beautiful: after an overview (chapter 3), Allison analyzes each of the "four moments" of judgments of taste, namely their characterization as to quality, quantity, relation, and modality (chapters 4 to 7). He then provides an interpretation and defense of Kant's Deduction of judgments of taste (chapter 8). To characterize the difference of purpose between the exposition of the four moments and the Deduction, Allison borrows the first *Critique*'s distinction between *quid facti* and *quid juris* questions. The four moments are supposed to answer the former (*quid facti*), the Deduction is supposed to answer the latter (*quid juris*). I am not sure Allison is right to apply that distinction here. In the first *Critique*, Kant calls "questions of fact" (*quid facti*) those questions that concern the way empirical concepts are acquired. In the case of empirical concepts, he says, answering questions of fact is also answering questions of right, namely ex-

[1] *Kant's Transcendental Idealism* (New Haven: Yale, 1983); *Kant's Theory of Freedom* (New York: Cambridge, 1990); see also *Idealism and Freedom* (New York: Cambridge, 1996).

0022-362X/03/0009/487–92 © 2003 The Journal of Philosophy, Inc.

plaining why it is legitimate to apply such concepts to objects.[2] On the contrary, in the case of categories or a priori concepts, showing how they have been empirically acquired is *neither* exhausting the explanation of their origin *nor* legitimating their use in knowledge of objects. The a priori origin of the categories in logical functions of judgment is argued for in the Metaphysical Deduction of the Categories, which is thus a *justification of the claim* that categories have an origin that is both a priori and intellectual. Their applicability to objects of nature is justified in the Transcendental Deduction. Now Allison suggests that the distinction between the four moments and the Deduction, in the *Critique of the Power of Judgment*, parallels that between the Metaphysical and the Transcendental Deduction, in the *Critique of Pure Reason*. If he is correct, as I think he is, then the four moments cannot possibly be a mere answer to a "*quid facti*" question, in the first *Critique*'s sense. Just like the Metaphysical Deduction, they too are *the justification of a claim*: the claim that aesthetic judgments have an a priori origin. The role of the Deduction of judgments of taste, parallel to that of the Transcendental Deduction in the first *Critique*, is then to justify their claim to universal and necessary agreement on the part of all judging subjects.

Of course, Allison could respond by defending his right to give the distinction between *quid facti* and *quid juris* a different sense than the sense Kant gives it in the first *Critique*. But I think we do not have just a mere quarrel of words here. I suggest that recognizing the justificatory aspect of the four moments, parallel to that of the Metaphysical Deduction in the first *Critique*, helps resolve an issue that Allison himself addresses at length: that of the relationship between the fourth moment (modality: the subjective necessity of judgments of taste) and the Deduction proper. I will say more on this shortly.

Unlike most previous commentators, Allison takes quite seriously Kant's use of the table of the logical functions of judgment expounded in the first *Critique* as a leading thread in his exposition of the four moments of judgments of taste (namely, Kant's analysis of those judgments according to the four titles of the table of logical functions). As Allison and others have noted, however, here Kant starts with quality, not quantity. This is because the first question to be answered is: *What* is being asserted in a judgment of taste, that is, what is the meaning of the predicate 'beautiful' (or as the case may be, 'ugly')? Kant's response: what is being asserted is the object's being the occasion of a peculiar kind of "pleasure" in us. Allison provides a carefully

[2] *Critique of Pure Reason*, A87/B119.

crafted analysis of what Kant calls the "disinterested" character of the feeling of pleasure expressed in judgments of taste. He defends the striking thesis that this feeling has an intentional aspect: it is not only *caused* by the free play of faculties in us, but it is also precisely the state of mind by way of which we become *aware* of this free play. One might object to Allison that the "free play of faculties" is not a state of mind we are *aware* of. Rather, it is a state the transcendental philosopher offers as *an explanation* for the peculiarity of the pleasure experienced and expressed in a judgment of taste.

According to Kant, the second moment (the "subjective universality" of judgments of taste) rests on the first: *because* the feeling of pleasure expressed in the predicate is disinterested (first moment), the judgment can make a claim to be shared by all (second moment). Allison deploys a great deal of care and exegetical subtlety in defending Kant's view of the relation between distinterestedness and subjective universality of judgments of taste against standard objections. I did not find the defense convincing, and I wonder whether it is all that necessary. Is it that important to Kant's analysis that the second moment be justified by the first? Why would it not simply describe a second characteristic of judgments of taste: they make a claim to the agreement of all judging subjects, so that we have here one more ground for maintaining that *if* there are indeed such judgments, then they must have an a priori origin? Of course, in proposing such a reading I am also proposing to interpret Kant's use of his leading thread in a different way than Allison does. Allison reads the four moments as a progressive argument in which each step depends on the preceding ones. I propose to read the four moments as a checklist as it were, where each title describes a specific aspect of any judgment (like it does in the table of the first *Critique*) and thus helps elucidate the nature of judgments of taste (what is being asserted of what (quality and quantity), and on what grounds (relation)?).

Allison notes that in the third moment (relation), Kant departs from the meaning of the corresponding title in the table of the first *Critique* (119). I find this claim surprising. A judgment of taste is easily recognized as a categorical judgment, so that the question to be answered is: What, in the logical subject of the judgment ('this *X* is beautiful') grounds the attribution of the predicate? Kant's answer: it is the "form" of the object, "as apprehended" by the judging subject. In this connection, Allison provides an illuminating analysis of what Kant calls the "formal purposiveness" and "purposiveness of form" of the object judged to be beautiful (125–38).

Allison's analysis of the fourth moment (the "subjective necessity" of judgments of taste) and of the Deduction that immediately follows

is particularly impressive. He makes two important claims about the fourth moment: one concerning its relation to the other three, the other concerning its relation to the Deduction. (1) The moment of modality, Allison maintains, does not add a new element to our understanding of the form and content of judgments of taste. Rather, it unifies the three previous moments and derives from this unifying focus a characterization of the evaluative or normative force of those judgments (see 144). (2) In doing so, the fourth moment prepares the ground for the Deduction without overstepping its role. In the fourth moment, Kant characterizes aesthetic judgments as making a claim to subjective necessity (we claim that others ought necessarily to agree with our judgment, without having any objective ground for our claim). Kant then points out the existence of a universally shared and sharable agreement of imagination and understanding ("common sense") as the ground for all *cognitive* judgments.[3] This prepares the ground for the argument of the Deduction (§38), according to which an agreement of imagination and understanding *which, unlike their agreement in cognitive judgments, is not bound by rules*, grounds the claim that aesthetic judgments can be, and ought to be universally shared. So, as I indicated earlier in this review, the fourth moment does have *some* justificatory role: it justifies the claim that aesthetic judgments, because of their claim to subjective necessity, have to have an a priori ground. It also points the way to the nature of this a priori ground. But only the Deduction specified the nature of that ground and thus justifies (or so Kant thinks, and Allison approves) to the universal and necessary subjective validity of pure judgments of taste.

I am not sure Allison is entirely successful in his defense of the Deduction. As other commentators have pointed out, it still seems that Kant has at most proved the *possibility* of obtaining other judging subjects' agreement to my aesthetic judgment. He has not proved the *necessity* of a *universal* agreement, even under the supposition that my judgment is formed under the subjective norm of the *sensus communis aestheticus*, that is, the universally sharable harmonious play of imagination and understanding. In order to move from the justification of that mere *possibility* to that of the *necessity* of a *universal* agreement, it seems that one needs to bring into play the normative request carried by the idea of a "universal voice," the voice of a universal community of judging subjects. Such a request is itself not sufficiently grounded in the mere play of imagination and understanding in their free harmony. It has a source of its own, which the solution to the dialectic

[3] *Critique of the Power of Judgment*, §21.

will reveal to be the intelligible nature of judging subjects. And it grounds the "intellectual interest in the beautiful" stated and expounded in §42, immediately after the Deduction of judgments of taste.

This leads us to the topics addressed in part III, "The Moral and Systematic Significance of Taste." In chapter 9, Allison takes us back to the two Introductions to the *Critique of the Power of Judgment*, in order to review Kant's systematic account of the relation between taste and morality. In chapter 10, he analyzes the role of the "disinterestedness" of judgments of taste in educating us for morality. But his most striking contribution comes with chapter 11, where he analyzes Kant's solution to the Antinomy of judgments of taste. According to Allison, the "indeterminable concept" or "concept of the supersensible" which, according to Kant, provides its solution to the Antinomy of the Critique of Taste, is that...of the beautiful itself. Among the arguments Allison gives to support his thesis is the definition of the beautiful in the third moment of the Analytic of judgments of taste: "Beauty is the form of *purposiveness* insofar as it is perceived in the object *without the presentation of a purpose*." Purposiveness, Allison notes, is according to Kant a concept of reason. So, this characterization of the beautiful makes it a concept, or more precisely an idea, of reason, and so a good candidate for being the indeterminable concept of the supersensible that grounds Kant's solution to the Antinomy of the Critique of Taste. Moreover, the concept of 'beautiful' is doubly indeterminable, since it is defined as the form of purposiveness *without the representation of a purpose*. And finally, in Kant's classification of representations,[4] there is nothing else for the representation 'beautiful' to be than an idea of reason, since it is certainly not a concept of the understanding, and it is not an intuition.

Now, although it is true that defining the beautiful as the "form of purposiveness without the representation of a purpose" leads to the idea of a supersensible ground for this purposiveness, this does not make the concept of the beautiful itself a concept of the supersensible. As Allison himself notes, the beautiful is certainly *sensible*. Moreover, although Allison is again correct in noting that it would be implausible to consider the beautiful as a concept of the understanding, taking it to be an idea of reason is perhaps not the only remaining option (*pace* Allison, 249). Why would it not be the case that the Critique of Aesthetic Judgment forces us to consider another kind of concept, a concept we might describe as a concept *of the merely reflective power of*

[4] Cf. *Critique of Pure Reason*, A319–20/B376–77

judgment? Compare with what Kant says of the concept of purposiveness in the second part of the *Critique of the Power of Judgment*, the *Critique of Teleological Judgment*: in its use for our cognition of organisms, purposiveness is an idea of reason *for reflective judgment*.[5] Here the concept of the beautiful might be a concept *of reflective judgment* itself. This concept does contain, among other marks, an idea of reason for reflective judgment, that of purposiveness. But it also has other marks, according to the other "moments" of the Analytic of the Beautiful: it is "the object of a satisfaction devoid of all interest." It is "what, without a concept, is liked universally." It is "what, without a concept, is cognized as the object of a necessary satisfaction." To these characterizations, the solution to the Antinomy adds the thesis that full due can be given to the paradoxical character of the concept of the beautiful only by taking into account the *indeterminable concept of the supersensible* that is the *ground* of the sensible object judged to be beautiful, as well as of the judging subjects themselves. Allison's analyses help us better understand these points. I do not think they justify characterizing the concept of the beautiful *itself* as being the indeterminable concept of the supersensible that, according to Kant, provides the solution to the Antinomy.

Part IV further illuminates the relation between the sensible and the supersensible in Kant's explanation of beauty in nature and art, by considering what Allison calls two "parerga" of Kant's theory of taste: Kant's theory of fine art and genius (chapter 12) and Kant's theory of the sublime (chapter 13).

Because of its detailed and subtle textual analyses, and because of the novelty and boldness of its core theses—about Kant's method in the *Critique of the Power of Judgment*, about the respective roles of the "four moments" and the Deduction of Taste, about the Antinomy and its solution, about the relation between art and morality—Allison's new book will remain an indispensable reference for any discussion of Kant's theory of taste.

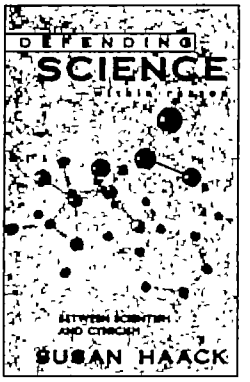
BÉATRICE LONGUENESSE

Princeton University

[5] *Critique of the Power of Judgment* §75.

Prometheus Books

DEFENDING SCIENCE—WITHIN REASON



Between Scientism and Cynicism
Susan Haack

"Reading Susan Haack's book is to witness a rarity: the level-headed appraisal of current issues and ideas, seen through the eyes of a widely informed and sensible philosopher who explores the middle ground between competing extremes, providing the guidance of a civilized mind."
—Gerald Holton, Mallinckrodt Professor of Physics and History of Science, Emeritus, Harvard University

"... original work from a scholar of world reputation on the nature of science and its impact upon cultural issues that engage the public interest...deep and accessible, with arguments offered in a signature, honest prose of great clarity..."
—Paul R. Gross, former Provost and University Professor of Life Sciences, University of Virginia

"... [a] comprehensive defense of science ... appropriately argued without resort to clichés—pleasurable prose."
—Jacques Barzun, former Provost and Seth Low Prof. of History, Columbia University; author of *From Dawn to Decadence*

"[Susan Haack is] one of those rare contemporary philosophers I can read with pleasure." —Steven Weinberg, Nobel Laureate in Physics, from *Facing Up: Science and Its Cultural Adversaries*

390 pp • ISBN 1-59102-117-0 • HC $28

Humanity Books

NATURALIZING THE TRANSCENDENTAL

A Pragmatic View
Sami Pihlström

In this powerful and original work, Finnish philosopher Sami Pihlström argues that the transcendental and the pragmatist traditions should converge, or at least supplement each other, instead of being regarded as rivals. He proposes a workable middle ground between extreme realism on the one hand and extreme postmodernist skepticism and relativism on the other.

330 pp • ISBN 1-59102-104-9 • HC $49

A HOUSE DIVIDED

Comparing Analytic and Continental Philosophy
Edited by C. G. Prado

The aim of this collection is to reconsider the often facile characterization of major thinkers as belonging to either one or the other philosophical tradition.

350 pp • ISBN 1-59102-105-7 • PB $35

For exam copies contact LRisio@prometheusbooks.com

Call toll free (800) 421-0351 • Fax (716) 691-0137 • www.prometheusbooks.com

New from
Oxford







Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556
Visit our website at www.oup-usa.org

OXFORD

Moral Realism
A Defence
RUSS SHAFER-LANDAU
Russ Shafer-Landau argues that there are moral principles that are true independently of what anyone, anywhere, happens to think of them. His central thesis, as well as the many novel supporting arguments used to defend it, will spark much controversy among those concerned with the foundations of ethics.
2003 $45.00

Libertarian Accounts of Free Will
RANDOLPH CLARKE
Bringing to bear recent work on action, causation, and causal explanation, Clarke defends a type of event-causal view—one in which a free action must be non-deterministically caused by its immediate causal antecedents—from the charges concerning rationality and diminished control. Clarke subtly explores the extent to which event-causal accounts can secure the things for the sake of which we value free will, judging their success here to be limited.
September 2003 $49.95

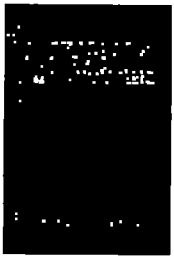
New in Paperback!
Human Nature and the Limits of Science
Edited by JOHN DUPRE
"The interesting question... is why today's physicalist pathology should have achieved a considerable cultural plausibility, and it is here that we find this book's most innovative and important contribution to the philosophy of science."—David Hawkes, *Times Literary Supplement*
October 2003 paper $16.95

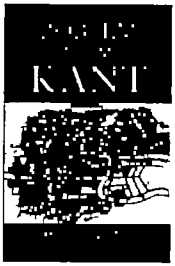
The Oxford Handbook of Metaphysics
MICHAEL J. LOUX and DEAN W. ZIMMERMAN
The Oxford Handbook of Metaphysics offers the most authoritative and compelling guide to this diverse and fertile field of philosophy. Twenty-four of the world's most distinguished specialists provide brand-new essays about 'what there is': what kinds of things there are, and what relations hold among entities falling under various categories. They give the latest word on such topics as identity, modality, time, causation, persons and minds, freedom, and vagueness.
September 2003 $99.00

Making Things Happen
A Theory of Causal Explanation
JAMES WOODWARD
Woodward's long awaited book is an attempt to construct a comprehensive account of causation explanation that applies to a wide variety of causal and explanatory claims in different areas of science and everyday life. The book engages some of the relevant literature from other disciplines, as Woodward weaves together examples, counterexamples, criticisms, defenses, objections, and replies into a convincing defense of the core of his theory, which is that we can analyze causation by appeal to the notion of manipulation.
September 2003 $55.00

New from
Oxford







Prices are subject to change and apply only in the US To order, or for more information, please call 1-800-451-7556. Visit our website at www.oup-usa.org

OXFORD

The Things We Mean
STEPHEN SCHIFFER
Stephen Schiffer presents a groundbreaking account of meaning and belief, and shows how it can illuminate a range of crucial problems regarding language, mind, knowledge, and ontology. He introduces the new doctrine of 'pleonastic propositions' to explain what the things we mean and believe are. He discusses the relation between semantic and psychological facts, on the one hand, and physical facts, on the other; vagueness and indeterminacy, moral truth; conditionals; and the role of propositional content in information acquisition and explanation
October 2003 paper $24.95 cloth $65.00

New in Paperback!
Morals from Motives
MICHAEL SLOTE
"Slote here develops a *sentimentalist* approach to virtue ethics, one that is inspired by the Humean rather than the Aristotelian tradition in ethics. This book should be required reading for those interested in virtue ethics and understanding how it compares to other moral theories."—Julia Driver, *Journal of Ethics*
paper August 2003 $19.95

New in Paperback!
Problems from Kant
JAMES VAN CLEVE
"There can be few recent books on the problems of Kant's *First Critique* that treat so great a range of arguments with such seriousness and sophistication. In their precision, originality and brevity, these are gems of analysis. This is a splendid book."
—*The Philosophical Review*
paper September 2003 $24.95

Science and Partial Truth
A Unitary Approach to Models and Scientific Reasoning
NEWTON C. A. DA COSTA and STEVEN FRENCH
Da Costa and French explore the consequences of adopting a "pragmatic" notion of truth in the philosophy of science—in other words, accepting a theory as valid when it may only be *partially* true rather than *wholly* true. Their framework sheds new light on issues dealing with belief, theory acceptance, and the realism-antirealism debate, as well as the nature of scientific models and their heuristic development
August 2003 $49.95

New in Paperback!
Natural Goodness
PHILIPPA FOOT
"*Natural Goodness* is an exciting and provoking book, more interesting than most books in moral philosophy. What she has given us both in this book and elsewhere deserves nothing but intellectual and moral gratitude."—Alasdair MacIntyre, *Philosophical Quarterly*
paper October 2003 $17.95

OUTSTANDING SCHOLARSHIP

The Evolution of Reason
Logic as a Branch of Biology
William S. Cooper
Cambridge Studies in Philosophy and Biology
0-521-54025-9 Paperback $28.00*

An Introduction to the Philosophy of Art
Richard Eldridge
0-521-80135-4 Hardback $60.00*
0-521-80521-X Paperback $19.00*

The Cambridge Companion to Medieval Jewish Philosophy
Edited by Daniel H. Frank and Oliver Leaman
Cambridge Companions to Philosophy
0-521-65207-3 Hardback $60.00*
0-521-65574-9 Paperback $24.00*

Richard Rorty
Edited by Charles Guignon and David R. Hiley
Contemporary Philosophy in Focus
0-521-80058-7 Hardback $55.00
0-521-80489-2 Paperback $20.00

The Political Philosophy of Needs
Lawrence Hamilton
0-521-82782-5 Hardback $60.00

Iconoclasm in Aesthetics
Michael Kelly
0-521-82209-2 Hardback $60.00

Novalis: *Fichte Studies*
Edited by Jane Kneller
Cambridge Texts in the History of Philosophy
0-521-64353-8 Hardback $58.00
0-521-64392-9 Paperback $22.00



Available in bookstores or from

CAMBRIDGE UNIVERSITY PRESS 800-872-7423 us.cambridge.org/philosophy

FROM CAMBRIDGE

Faking It
William Ian Miller
0-521-83018-4 Hardback $29.00

Biological Complexity and Integrative Pluralism
Sandra D. Mitchell
Cambridge Studies in Philosophy and Biology
0-521-81753-6 Hardback $70.00*
0-521-52079-7 Paperback $26.00*

The Just War Revisited
Oliver O'Donovan
Current Issues in Theology 2
0-521-83138-5 Hardback $45.00*
0-521-53899-8 Paperback $17.00*

Barbarism and Religion
Volume 3: The First Decline and Fall
J.G.A. Pocock
0-521-82445-1 Hardback $60.00

John Searle
Edited by Barry Smith
Contemporary Philosophy in Focus
0-521-79288-6 Hardback $65.00
0-521-79704-7 Paperback $20.00

Seneca: *De Otio; De Brevitate Vitae*
Edited by Gareth Williams
Cambridge Greek and Latin Classics
0-521-58223-7 Hardback $70.00
0-521-58806-5 Paperback $24.00

Religion after Metaphysics
Edited by Mark Wrathall
0-521-82498-2 Hardback $55.00*
0-521-53196-9 Paperback $20.00*

Democratic Distributive Justice
Ross Zucker
0-521-53355-4 Paperback $23.00



Available in bookstores or from

**Prices subject to change.*

CAMBRIDGE UNIVERSITY PRESS 800-872-7423 us.cambridge.org/philosophy



THE BLANK SLATE
The Modern Denial of Human Nature
Steven Pinker

"[Pinker] makes his main argument persuasively and with great verve ... ought to be read by anybody who feels they have had enough of nature-nurture rows or who thinks they already know where they stand on the science wars....It could change their mind." —*The Economist.*

Penguin 528 pp 0-14-200334-4 $16.00

THE PORTABLE HANNAH ARENDT
Edited by Peter Baehr

The first general anthology of Arendt's writings, this volume includes: *The Origins of Totalitarianism, The Human Condition, and Eichmann in Jerusalem* as well as several other provocative essays and correspondence with influential figures of the time.

Penguin Classics 576 pp 0-14-243766-5 $17.00

THE HEART OF PHILOSOPHY
Jacob Needleman

Needleman's celebrated introduction to philosophy as a living art, with a new preface by the author. "His work now becomes a way station [for] those who are seriously exploring the transformation of consciousness." —Theodore Roszak, *Los Angeles Times.*

Tarcher 240 pp 1-585542-251-7 $14.95

HAMLET: Poem Unlimited
Harold Bloom

"Far superior to existing theories of performance and worth yards of criticism for each well-wrought page...will delight everyone." —*Library Journal.*

Riverhead 192 pp 1-57322-233-X $19.95

THE GREEK SOPHISTS
Translated by John Dillon & Tania Gergel
Introduction by John Dillon

In this new translation each chapter is organized around the work of one character, including Gorgias, Prodicus, Protagoras, and Antiphon, among others.

Penguin Classics 288 pp 0-14-043669-8 $14.00

THE ART OF WAR
Sun-tzu
Edited, Translated, and with an Introduction by John Minford

"This fundamental volume could be one of the best in recent years." —*Publishers Weekly.*

Penguin Classics 384 pp 0-14-043919-8 $14.00

FRAGMENTS
The Collected Wisdom of Heraclitus
Heraclitus
Translated by Brooks Haxton
Foreword by James Hillman

"A pellucid and informed translation." —Rita Dove, *The Washington Post.*

Penguin Classics 128 pp 0-14-243786-4 $13.00

PENGUIN GROUP (USA)
Academic Marketing Department, 375 Hudson Street, New York, NY 10014

THE JOURNAL OF PHILOSOPHY

Calcutta University Library
ALIPORE CAMPUS

VOLUME C, NUMBER 10
OCTOBER 2003

page

493 *Moving beyond Metaphors: Understanding the Mind for What It Is* Chris Eliasmith

521 *The End of Counterpart Theory* Trenton Merricks

550 NOTES AND NEWS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX (212) 932-8721.

You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2003

SUBSCRIPTIONS (12 issues)

Individuals	$85.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset and printed by Capital City Press, Montpelier, VT.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices

POSTMASTER: Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027

© Copyright 2003 by the Journal of Philosophy, Inc.

ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY
VOLUME C, NO. 10, OCTOBER 2003

Calcutta University Central Library CAMPUS

MOVING BEYOND METAPHORS: UNDERSTANDING THE MIND FOR WHAT IT IS*

In the last fifty years, there have been three major approaches to understanding cognitive systems and theorizing about the nature of the mind: *symbolicism*, *connectionism*, and *dynamicism*. Each of these approaches has relied heavily on a preferred metaphor for understanding the mind. Most famously, symbolicism, or classical cognitive science, relies on the "mind as computer" metaphor. Under this view, the mind is the software of the brain. Jerry Fodor,[1] for one, has argued that the impressive theoretical power provided by this metaphor is good reason to suppose that cognitive systems have a symbolic "language of thought" which, like a computer programming language, expresses the rules that the system follows. Fodor claims that this metaphor is essential for providing a useful account of how the mind works.

Similarly, connectionists have relied on a metaphor for providing their account of how the mind works. This metaphor, however, is much more subtle than the symbolicist one; connectionists presume that the functioning of the mind is like the functioning of the brain. The subtlety of the "mind as brain" metaphor lies in the fact that connectionists, like symbolicists, are materialists. That is, they also hold that the mind is the brain. When providing psychological descriptions, however, it is the metaphor that matters, not the identity. In deference to the metaphor, the founders of this approach call it

* Special thanks to Charles H. Anderson. Thanks as well to William Bechtel, Ned Block, David Byrd, Rob Cummins, Brian Keeley, Brian McLaughlin, William Ramsey, Paul Thagard, and Charles Wallis for comments on earlier versions. Funding has been provided in part by the Mathers Foundation, the McDonnell Center for Higher Brain Function, and the McDonnell Project for Philosophy and the Neurosciences.

[1] *The Language of Thought* (New York: Crowell, 1975).

0022-362X/03/0010/493–520 © 2003 The Journal of Philosophy, Inc.

"brain-style" processing, and claim to be discussing "abstract networks."[2] This is not surprising since the computational and representational properties of the nodes in connectionist networks bear little resemblance to neurons in real biological neural networks.[3]

Proponents of dynamicism also rely heavily on a metaphor to understand cognitive systems. Most explicitly, Tim van Gelder[4] employs the Watt Governor as a metaphor for mind. It is through his analysis of the best way to characterize this dynamic system that he argues that cognitive systems, too, should be understood as nonrepresentational, low-dimensional, dynamical systems. Like the Watt Governor, van Gelder argues, cognitive systems are essentially dynamic and can only be properly understood by characterizing their state changes through time. The "mind as Watt Governor" metaphor suggests that trying to impose any kind of discreteness, either temporal or representational, will lead to a mischaracterization of minds.

Notably, each of symbolicism, connectionism, and dynamicism rely on metaphor not only for explanatory purposes, but also for developing their conceptual foundations in understanding the target of the metaphor; that is, the mind. For symbolicists, the properties of Turing machines become shared with minds. For connectionists, the character of representation changes dramatically. Mental representations are taken to consist of "sub-symbols" associated with each node, while "whole" representations are real-valued vectors in a high-dimensional property space.[5] Finally, for the dynamicists, because the Watt Governor is best described by dynamic systems theory, which makes no reference to computation or representation, our theories of mind need not appeal to computation or representation either.

In this article, I want to suggest that it is time to move beyond these metaphors. We are in the position, I think, to understand the mind for what it is: the result of the dynamics of a complex, physical, information processing system, namely the brain. Clearly, in some ways this is a rather boring thesis to defend. It is just a statement of plain old "monistic materialism" or "token identity theory," call it

[2] James L. McClelland and David E. Rumelhart, "Future Directions," in McClelland and Rumelhart, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Volume 2 (Cambridge: MIT, 1986), pp. 547–52.

[3] See Chapter 10 of William Bechtel and Adele Abrahamsen, *Connectionism and the Mind: Parallel Processing, Dynamics, and Evolution in Networks*, Second Edition (Malden, MA: Blackwell, 2001).

[4] "What Might Cognition Be, If Not Computation?" this JOURNAL, XCI, 7 (July 1995): 345–81.

[5] See, for example, Paul Smolensky's "On the Proper Treatment of Connectionism," *Behavioral and Brain Sciences*, XI, 1 (1988): 1–23.

what you will. It is, in essence, just the uncontroversial view that, so far as we know, you do not have a mind without a brain. But, I further want to argue that the best way to understand this physical system is by using a different set of conceptual tools than those employed by symbolicists, connectionists, and dynamicists individually. That is, the right toolbox will consist in an extended subset of the tools suggested by these various metaphors.

The reason we need to move beyond metaphors is because, in science, analogical thinking can sometimes constrain available hypotheses. This is not to deny that analogies are incredibly useful tools at many points during the development of a scientific theory. It is only to say that, sometimes, analogies only go so far. Take, for instance, the development of the current theory of the nature of light. In the nineteenth century, light was understood in terms of two metaphors: light as a wave, and light as a particle. Thomas Young was the best known proponent of the first view, and Isaac Newton was the best known proponent of the second. Each used their favored analogy to suggest new experiments, and develop new predictions.[6] Thus, these analogies played a role similar to that played by the analogies discussed above in contemporary cognitive science. As we know in the case of light, however, both analogies are false. Hence the famed "wave-particle duality" of light: sometimes it behaves like a particle; and sometimes it behaves like a wave. Neither analogy by itself captures all the phenomena displayed by light, but both are extremely useful in characterizing some of those phenomena. So, understanding what light *is* required moving beyond the metaphors.

I want to suggest that the same is true in the case of cognition. Each of the metaphors mentioned above has some insight to offer regarding certain phenomena displayed by cognitive systems. However, none of these metaphors is likely to lead us to all of the right answers. Thus, my project in trying to move beyond these metaphors is a synthetic one. I want to provide a way of understanding cognitive systems that draws on the strengths of symbolicism, connectionism, and dynamicism. The best way of doing this is to understand minds for what they are. To phrase this as a conditional, if minds are the behavior of complex, dynamic, information processing systems, then we should use the conceptual tools that we have for understanding such systems when trying to understand minds. I outline here a general theory

[6] For a detailed description of the analogies, predictions, and experiments, see Chris Eliasmith and Paul Thagard, "Particles, Waves and Explanatory Coherence," *British Journal of the Philosophy of Science*, XLVIII (1997). 1–19.

that describes *representation* and *dynamics* in neural systems (*R&D theory*) that realizes the consequent of this conditional. I argue that R&D theory can help unify neural and psychological explanations of cognitive systems and that the theory suggests a need to re-evaluate standard functionalist claims.

First, however, it is instructive to see how R&D theory does not demand the invention of new conceptual tools; the relevant tools are already well tested. So, in some ways, the theory is neither risky nor surprising. What is surprising, perhaps, is that our most powerful tools for understanding the kinds of systems that minds are have yet to be applied to minds. I suggest that this surprising oversight is due to an overreliance on the "mind as computer" metaphor.

I. A BRIEF HISTORY OF COGNITIVE SCIENCE

While the main purpose of this article is clearly not historical, a brief perusal of the relevant historical landscape helps situate both the theory and the subsequent discussion.

I.1. Prehistory. While much is sometimes made of the difference between philosophical and psychological behaviorism, there was general agreement on this much: internal representations, states, and structures are irrelevant for understanding the behavior of cognitive systems. For psychologists, like John B. Watson and B.F. Skinner, this was true because only input/output relations are scientifically accessible. For philosophers, like Gilbert Ryle, this was true because mental predicates, if they were to be consistent with natural science, must be analyzable in terms of behavioral predicates. In either case, looking inside the "black box" that was the object of study, was prohibited for behaviorists.

Interestingly, engineers of the day respected a similar constraint. In order to understand dynamic physical systems, the central tool they employed was (classical) control theory. Classical control theory, notoriously, only characterizes physical systems in terms of their input/output relations in order to determine the relevant controller. Classical control theory was limited to designing nonoptimal, single-variable, static controllers and depended on graphical methods, rules of thumb, and did not allow for the inclusion of noise.[7] While the limitations of classical controllers and methods are now well known, they nevertheless allowed engineers to build systems of a kind they had not systematically built before: goal-directed systems.

[7] For a succinct description of the history of control theory, see Frank L. Lewis's *Applied Optimal Control and Estimation* (New York: Prentice-Hall, 1992).

While classical control theory was useful (especially in the 1940s) for building warhead guidance systems, some researchers thought it was clearly more than that. They suggested that classical control theory could provide a theoretical foundation for describing living systems as well. Most famously, the interdisciplinary movement founded in the early 1940s known as "cybernetics" was based on precisely this contention.[8] Cyberneticists claimed that living systems were also essentially goal-directed systems. Thus, closed-loop control, it was argued, should be a good way to understand the behavior of living systems. Given the nature of classical control theory, cyberneticists focused on characterizing the input/output behavior of living systems, not their internal processes. With the so-called "cognitive revolution" of the mid-1950s, interest in cybernetics waned due in part to its close association with, and similar theoretical commitments to, behaviorism.

I.2. The cognitive revolution. In the mid-1950s, with the publication of a series of seminal papers,[9] the "cognitive revolution" took place. One simplistic way to characterize this shift from behaviorism to cognitivism is that it became no longer taboo to look inside the black box. Quite the contrary: internal states, internal processes, and internal representations became standard fare when thinking about the mind. Making sense of the insides of that black box was heavily influenced by concurrent successes in building and programming computers to perform complex tasks. Thus, many early cognitive scientists saw, when they opened the lid of the box, a computer. Furthermore, as explored in detail by Fodor, "Computers show us how to connect semantical with causal properties for *symbols*." So computers have what it takes to be intentional minds.[10] Once cognitive scientists began to think of minds as computers, a number of new theoretical tools became available for characterizing cognition. For instance, the computer's theoretical counterpart, the Turing machine, suggested novel philosophical theses, including functionalism and multiple realizability, about the mind. More practically, the typical architecture of com-

[8] For a statement of the motivations of cybernetics, see Arturo Rosenblueth, Norbert Wiener, and Julian Bigelow, "Behavior, Purpose, and Teleology," *Philosophy of Science*, x (1943): 18–24.

[9] These papers include, but are not limited to: A. Newell, C. Shaw, and H. Simon, "Elements of a Theory of Human Problem Solving," *Psychological Review*, LXV (1958): 151–66; G. Miller, "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information," *Psychological Review*, LXIII (1956). 81–97; Jerome S. Bruner, Jacqueline Goodnow, and George Austin, *A Study Of Thinking* (New York: Wiley, 1956).

[10] *Psychosemantics* (Cambridge: MIT, 1987), p. 18.

puters, the von Neumann architecture, was thought by many to be relevant for understanding our cognitive architecture.

Adoption of the von Neumann architecture for understanding minds was seen by many as poorly motivated, however. As a result, the early 1980s saw a revival of the so-called "connectionist" research program. Rather than adopting the architecture of a digital computer, these researchers felt that an architecture more like that seen in the brain would provide a better model for cognition.[11] As a result of this theoretical shift, connectionists were very successful at building models sensitive to statistical structure, and could begin to explain many phenomena not easily captured by symbolicists (for example, object recognition, generalization, and learning).

For some, however, connectionists had clearly not escaped the influence of the "mind as computer" metaphor. Connectionists still spoke of representations, and thought of the mind as a kind of computer. These critics argued that minds are not essentially computational, they are essentially physical, dynamic systems.[12] They suggested that if we want to know which functions a system can actually perform in the real world, we must know how to characterize the system's dynamics. Furthermore, since cognitive systems evolved in dynamic environments, we should expect evolved control systems, like brains, to be more like the Watt Governor—dynamic, continuous, coupled directly to what they control—than like a discrete state Turing machine that computes over "disconnected" representations. As a result, these "dynamicists" suggested that dynamic systems theory, not computational theory, was the right quantitative tool for understanding minds. They claimed that notions like 'chaos', 'hysteresis', 'attractors', and 'state-space' underwrite the conceptual tools best suited for describing cognitive systems.

I.3. A puzzling oversight. In some ways, dynamicists revived the commitments of the predecessors of the cognitive revolution. Notably, the Watt Governor is a standard example of a classical control system. If minds are to be like Watt Governors, they are to be like classical control systems; just what the cyberneticists had argued. One worry with this retrospective approach is that the original problems come

[11] As discussed in both Smolensky and the introduction to Rumelhart and McClelland, eds. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Volume 1 (Cambridge: MIT, 1986).

[12] See the various contributions to Robert F. Port and van Gelder, eds., *Mind as Motion: Explorations in the Dynamics of Cognition* (Cambridge: MIT, 1995), especially the editors' introduction.

along with the original solutions. The limitations of classical control theory are severe, so severe that they will probably not allow us to understand a system as complex as the brain.

An important series of theoretical advances in control theory went completely unnoticed during the cognitive revolution, however. During the heyday of the computer, in the 1960s, many of the limitations of classical control theory were rectified with the introduction of what is now known as "modern" control theory.[13] Modern control theory introduced the notion of an "internal system description" to control theory. An internal system description is one that includes *system state variables* (that is, variables describing the state of the system itself) as part of the description (see Figure 1). It is interesting that with the cognitive revolution, researchers interested in the behavior of living systems realized they needed to "look inside" the systems they were studying and, at about the same time, researchers interested in controlling engineered systems began to "look inside" as well. Both, nearly simultaneously, opened their black box. As already discussed, however, those interested in cognitive behavior adopted the computer as a metaphor for the workings of the mind. Unfortunately, the ubiquity of this metaphor has served to distance the cognitive sciences from modern control theory. Nevertheless, I argue below that modern control theory offers tools better suited than computational theory for understanding biological systems as fundamentally physical, dynamic systems operating in changing, uncertain environments.

This is not to suggest that each of the dominant metaphors should be taken as irrelevant to our understanding of minds. Both connectionism and dynamicism highlight important limitations of the original "mind as computer" metaphor. Connectionism challenged the symbolicist conception of representation, noting how important statistical considerations are for capturing certain kinds of cognitive phenomena. Dynamicist critiques of symbolicism focused on its lack of a principled account of the temporal properties of cognitive systems.[14] Nevertheless, it was the symbolicists, armed with their metaphor, who rightly justified opening the black box. Furthermore, both connectionism and dynamicism introduced their own misleading metaphors.

[13] This introduction is largely credited to R Kalman in his "A New Approach to Linear Filtering and Prediction Problems," *ASME Journal of Basic Engineering*, LXXXII (1960): 35–45.

[14] van Gelder, "What Might Cognition Be, If Not Computation?"

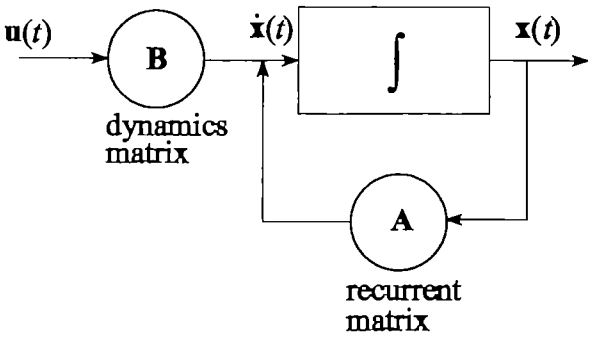


Figure 1: A modern control theoretic system description. The vector $\mathbf{u}(t)$ is the input to the system. **A** and **B** are matrices that define the behavior of the system, $\mathbf{x}(t)$ is the system state variable (generally a vector), and $\dot{\mathbf{x}}$ is the derivative of the state vector. The standard transfer function in control theory, as shown in the rectangle, is integration.[15]

II. REPRESENTATION AND DYNAMICS IN NEURAL SYSTEMS: A THEORY

Moving beyond metaphors—that is, taking seriously the view that minds are complex, physical, dynamic, and information processing systems—means using our best tools for describing systems with these properties. In the remainder of this section, I propose and defend a theory of representation and dynamics in neural systems (R&D theory) that takes precisely this approach. R&D theory relies on modern control theory, information theory, and recent results from neuroscience to provide an account of what minds are.[16]

Below I have broken this account into three parts. The first defines representation, the second describes computation, and the third section, on dynamics, shows how the preceding characterizations of representation and computation can be merged with control theory to provide an account of neural and cognitive function. The result, I argue, is a theory that avoids the weaknesses and capitalizes on the strengths of past approaches.

II.1. Representation. A central tenet of R&D theory is that we can adapt the information theoretic account of *codes* to understanding

[15] I have simplified this diagram for a generic linear system from the typical, truly general one found in most control theory texts by excluding the feedthrough and output matrices. Nothing turns on this simplification in this context.

[16] For an in-depth technical description of this approach, see Eliasmith and Charles H. Anderson, *Neural Engineering: Computation, Representation and Dynamics in Neurobiological Systems* (Cambridge: MIT, 2003).

representation in neural systems. Codes, in engineering, are defined in terms of a complimentary encoding and decoding procedure between two alphabets. Morse code, for example, is defined by the one-to-one relation between letters of the Roman alphabet, and the alphabet composed of a standard set of dashes and dots. The encoding procedure is the mapping from the Roman alphabet to the Morse code alphabet and the decoding procedure is its inverse.

In order to characterize representation in a cognitive/neural system, we can identify each of these procedures and their relevant alphabets. The encoding procedure is quite easy to identify: it is the transduction of stimuli by the system resulting in a series of neural "action potentials," or "spikes." The precise nature of this encoding has been explored in depth via quantitative models.[17] So, encoding is what neuroscientists typically talk about. When we show a cognitive system a stimulus, some neurons or other "fire." Unfortunately, neuroscientists often stop here in their characterization of representation, but this is insufficient. We also need to identify a decoding procedure—otherwise, there is no way to determine the relevance of the encoding for the system. If no information about the stimulus can be extracted from a spiking neuron, then it makes no sense to say that it represents the stimulus. Representations, at a minimum, must potentially be able to "stand-in" for their referents.

Quite surprisingly, despite typically nonlinear encoding, a good linear decoding can be found.[18] And, there are several established methods for determining linear decoders given the statistics of the neural populations that respond to certain stimuli.[19] Notably, these decoders are sensitive both to the temporal statistics of the stimuli and to what other elements in the population encode. Thus, if you have multiple neurons involved in the (distributed) representation of a time-varying object, they can "cooperate" to provide a better representation.

Having specified the encoding and decoding procedures, we still need to specify the relevant alphabets. While the specific cases will diverge greatly, we can describe the alphabets generally: neural responses (encoded alphabet) code physical properties (decoded alphabet). In fact, it is possible to be a bit more specific. Neuroscientists

[17] See James M. Bower and David Beeman, *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System* (Berlin: Springer, 1998) for a review of such models.

[18] As demonstrated by Fred Rieke et al., *Spikes: Exploring the Neural Code* (Cambridge: MIT, 1997), pp. 76–87.

[19] As discussed in Eliasmith and Anderson

generally agree that the basic element of the neural alphabet is the neural spike. There are many possibilities for how such spikes are used, however: average production rate of neural spikes (that is, a rate code); specific timings of neural spikes (that is, a timing code); population-wide groupings of neural spikes (that is, a population code); or the synchrony of neural spikes across neurons (that is, a synchrony code). Of these possibilities, arguably the best evidence exists for a combination of timing codes and population codes.[20] For this reason, let us take the combination of these basic coding schemes to define the alphabet of neural responses. Thus, the encoded alphabet is the set of temporally patterned neural spikes over populations of neurons.

It is much more difficult to be specific about the nature of the alphabet of physical properties. Of course, we can begin by looking to the physical sciences for categories of physical properties that might be encoded by nervous systems. Indeed, we find that many of the properties that physicists traditionally use do seem to be represented in nervous systems; for example, displacement, velocity, acceleration, wavelength, temperature, pressure, and mass. But there are many physical properties not discussed by physicists which also seem to be encoded in nervous systems; for example, red, hot, square, dangerous, edible, object, and conspecific. Presumably, all of these "higher-level" properties are inferred on the basis of representations of properties more like those that physicists talk about. In other words, encodings of 'edible' depend, in some complex way, on encodings of "low-level" physical properties like wavelength, velocity, and so forth. While R&D theory itself does not determine precisely what is involved in such complex relations, there is reason to suppose that R&D theory provides the necessary tools for describing such relations. To see why this is so, let us consider a simple example.

It is clearly important for an animal to be able to know where various objects in its environment are. As a result, in mammals, there are a number of internal representations of signals that convey and update this kind of information. One such representation is found in parietal areas, particularly in the lateral intraparietal cortex (LIP). For simplicity, let us consider the representation of only the horizontal position of an object in the environment. As a population, some neurons in this area *encode* an object's position over time. This representation can be understood as a scalar variable, whose units are

[20] For an overview of this evidence, see Rieke et al.; Eliasmith and Anderson; and L. Abbott, "Decoding Neuronal Firing and Modelling Neural Networks," *Quarterly Review of Biophysics*, XXVII, 3 (1994): 291–331.

degrees from midline (decoded alphabet), that is encoded into a series of neural spikes (encoded alphabet). Using the quantitative tools mentioned earlier, we can determine the relevant decoder. Once we have such a decoder, we can then estimate what the actual position of the object is given the neural spiking in this population. Thus we can determine precisely how well (or what aspects of) the original property (in this case, the actual position) is represented by the neural population. We can then use this characterization to understand the role that the representation plays in the cognitive system as a whole.

As mentioned, this is a simple example. But notice that it not only describes how to characterize representation, it also shows how we can move from talking about neurons to talking about "higher-level" variables, like object position. That is, we can move from discussing the "basic" representations (that is, neural spikes) to "higher-level" representations (that is, mathematical objects with units). This suggests that we can build up a kind of "representational hierarchy" that permits us to move further and further away from the neural-level description, while remaining responsible to it. For instance, we could talk about the larger population of neurons that encodes position in three dimensional space. We could dissect this higher-level description into its lower-level components (for example, horizontal, vertical, and depth positions), or we could dissect it into the activity of individual neurons. Which description we employ will depend on the kind of explanation we need. Notably, this hierarchy can be rigorously and generally defined to include scalars, vectors, functions, vector fields, and so forth.[21] The fact that all of the levels of such a hierarchy can be written in a standard form suggests that this characterization provides a unified way of understanding representation in neurobiological systems.

Note that the focus of this article is on how to characterize representational *states*, computations over these states, and the dynamics of these states. As a result, I do not directly address concerns related to content determination. This is largely because such a discussion would lead me far afield. Nevertheless, it is interesting to note that R&D theory is suggestive of a particular approach to content determination. Notice, first, that both the encoding and decoding are essential for determining the identity of a representation. This means that both what causes a neural state, and how that state is used by the system are likely to play a role in content determination. This suggests that some kind of two-factor theory is consistent with R&D theory. As well,

[21] This generalization is made explicit in Eliasmith and Anderson, pp 79–80.

a single neural state may play a role in multiple contents concurrently, because distinct, yet related, representations (and hence contents) can be identified at different levels of organization at the same time. This may initially seem problematic, but, because the relation between levels of representation is quantitatively defined (hence we know precisely what role a single neural state is playing in each of the representations defined over it), we should expect the parallel content relations to also be well defined. Of course, such comments only provide a hint of a theory of content, they do not constitute one.[22]

In any case, there is no reason to consider such a theory of content if its underlying theoretical assumptions are not appropriate to cognitive systems. So, we should notice that the strength of the previous characterization of representation lies in its generality. That is, regardless of what the higher-level representations look like (that is, what kind of mathematical objects with units they are), R&D theory will apply. So R&D theory, while having definite consequences for what constitutes a good representational story, is silent as to which particular one is correct for a given neural system. This is desirable for a theory of mind because higher-level representations are clearly theoretical postulates (at least at this point in the development of neuroscience). While we can directly measure the voltage changes of individual neurons, making claims about how they are grouped to represent the world is not easily confirmable. Presumably, the right representational story will be the most coherent and predictively successful one.

II.2. Computation. Of course, no representational characterization will be justified if it does not help us understand how the system functions. Luckily, a good characterization of neural representation paves the way for a good understanding of neural computation. This is because, like representations, computations can be characterized using decoding. But, rather than using the "representational decoder" discussed earlier, we can use a "transformational decoder." We can think of the transformational decoder as defining a kind of biased decoding. That is, in determining a transformation, we extract information *other than* what the population is taken to represent. The bias, then, is away from a "pure," or representational, decoding of the encoded information. For example, if we think that the quantity x is

[22] For instance, in order to understand the relation between representations at a given organizational level, we need to consider computational relations, as discussed in the next section. For an in-depth but preliminary discussion of a theory of content that is consistent with R&D theory, see my "How Neurons Mean A Neurocomputational Theory of Representational Content" (Ph.D. diss., Washington University, St. Louis, 2000).

encoded in some neural population, when defining the representation we determine the representational decoders that estimate x. When defining a computation, however, we identify transformational decoders that estimate some function, $f(x)$, of the represented quantity. In other words, we find decoders that, rather than extracting the signal represented by a population, extract some transformed version of that signal. The same techniques used to find representational decoders are applicable in this case, and result in decoders that can support both linear and nonlinear transformations.[25]

Given this understanding of neural computation, there is an important ambiguity that arises in the preceding characterization of representation. It stems from the fact that information encoded into a population may now be decoded in a variety of ways. Suppose we are again considering the population that encodes object position. Not surprisingly, we can decode that population to provide an estimate of object position. However, we can also decode that same information to provide an estimate of some function of object position (for example, the square). Since representation is defined in terms of encoding and decoding, it seems that we need a way to pick which of these possible decodings is the relevant one for defining the representation in the original population. To resolve this issue let me specify that what a population represents is determined by the decoding that results in the quantity that all other decodings are functions of. Thus, in this example, the population would be said to represent object position (since both object position and its square are decoded). Of course, object position is also a function of the square of object position (that is, $x = \sqrt{x^2}$). This further difficulty can be resolved by noticing that the right physical quantities (that is, the decoded alphabet) for representation are those that are part of a coherent, consistent, and useful theory. In other words, we characterize cognitive systems as representing positions because we characterize the world in terms of positions, and cognitive systems represent the world.

Importantly, this understanding of neural computation applies at all levels of the representational hierarchy, and accounts for complex transformations. So, for example, it can be used to define inference relations, traditionally thought necessary for characterizing the relations between high-level representations. Again consider the specific example of determining object position. Suppose that the available data from sensory receptors make it equally likely that an object is in one of two positions (represented as a bimodal probability distribution

[25] As demonstrated in Eliasmith and Anderson, pp. 143–60.

over possible positions). Also suppose, however, that prior information, in the form of a statistical model, favors one of those positions (perhaps one is consistent with past known locations given current velocity, and the other is not). Using the notion of computation defined above, it is straightforward to build a model that incorporates transformations between, and representations of (1) the top-down model, (2) the bottom-up data, and (3) the actual inferred position of the object (inferred based on Bayes's rule, for example). As expected, in this situation the most likely position given the a priori information would be the one consistent with the top-down model. If the bottom-up data is significantly stronger in favor of an alternate position, however, this will influence the preferred estimate, as expected.[24] So, although simple, performing linear decoding can support the kinds of complex transformations needed to articulate descriptions of cognitive behavior. Statistical inference is just one example.

Before moving on to a consideration of dynamics, it is important to realize that this way of characterizing representation and computation does not demand that there are "little decoders" inside the head. That is, this view does not entail that the system itself needs to decode the representations it employs. In fact, according to this account, there are no directly observable counterparts to the representational or transformational decoders. Rather, they are embedded in the synaptic weights between neighboring neurons. That is, coupling weights of neighboring neurons indirectly reflect a particular population decoder, but they are not identical to the population decoder. This is because connection weights are best characterized as determined by both the decoding of the incoming signal and the encoding of the outgoing signal. Practically speaking, this means that changing a connection weight both changes the transformation being performed and the tuning curve of the receiving neuron. As is well known from both connectionism and computational neuroscience, this is exactly what happens in such networks. In essence, the encoding/decoding distinction is not one that neurobiological systems need to respect in order to perform their functions, but it is extremely useful in trying to understand such systems and how they do, in fact, manage to perform those functions.

II.3. Dynamics. While it may be understandable that dynamics were

[24] For the technical details and results of the model described here, see Eliasmith and Anderson, pp. 275–83. For a brief discussion of more logic-like inference on symbolic representations, see section IV.1.

initially ignored by those studying cognitive systems as computational systems (theoretically, time is irrelevant for successful computation), it would be strange, indeed, to leave dynamics out of the study of minds as physical, neurobiological systems. Even the simplest nervous systems performing the simplest functions demand temporal characterizations (for example, locomotion, digestion, and sensing). It is not surprising, then, that single neural cells have almost always been modeled by neuroscientists as essentially dynamic systems. In contemporary neuroscience, electrophysiologists often analyze cellular responses in terms of 'onsets', 'latencies', 'stimulus intervals', 'steady states', 'decays', and so forth—these are all terms describing temporal properties of a neuron's response. The fact is, the systems under study in neurobiology are dynamic systems and as such they make it very difficult to ignore time.

Notably, modern control theory was developed precisely because understanding complex dynamics is essential for building something that works in the real world. Modern control theory permits both the analysis and synthesis of elaborate dynamic systems. Because of its general formulation, modern control theory applies to chemical, mechanical, electrical, digital, or analog systems. As well, it can be used to characterize nonlinear, time-varying, probabilistic, or noisy systems. As a result of this generality, modern control theory is applied to a huge variety of control problems, including autopilot design, spacecraft control, design of manufacturing facilities, robotics, chemical process control, electrical systems design, design of environmental regulators, and so on. It should not be surprising, then, that it proves useful for the analysis of the dynamics of cognitive, neurobiological systems as well.

Having identified quantitative tools for characterizing dynamics, and for characterizing representation and computation, how do we bring them together? An essential step in employing the techniques of control theory is identifying the system state variable ($\mathbf{x}(t)$ in Figure 1). Given the preceding analysis of representation, it is natural to suggest that the state variable *just is* the neural representation.

Things are not quite so simple, however. Because neurons have intrinsic dynamics dictated by their particular physical characteristics, we must adapt standard control theory to neurobiological systems. Fortunately, this can be done without loss of generality.[25] As well, all of the computations needed to implement such systems can be implemented using transformations as defined earlier. As a result, we can directly apply the myriad techniques for analyzing complex

[25] For the relevant derivations, see Eliasmith and Anderson, pp 221–25.

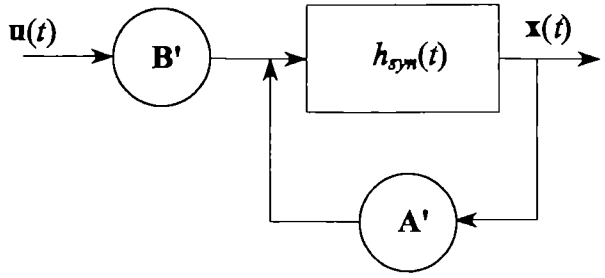


Figure 2: A control theoretic description of neurobiological systems. All variables are the same as in Figure 1. However, the matrices **A'** and **B'** take into account that there is a different transfer function, $h_{syn}(t)$, than in Figure 1. As well, $\mathbf{x}(t)$ is taken to be represented by a neural population.

dynamic systems that have been developed using modern control theory to this quantitative characterization of neurobiological systems.

To get a sense of how representation and dynamics can be integrated, let us revisit the simple example introduced previously: object position representation in area LIP. Note that animals often need to know not just where some object currently is, they need to remember where it was. Decades of experiments in LIP have shown that neurons in this area have sustained responses during the interval between a brief stimulus presentation and a delayed "go" signal.[25] In other words, these neurons seem to underlie the (short-term) memory of where an interesting object is in the world. Recall that I earlier characterized this area as representing $x(t)$, the position of an object. Now we know the dynamics of this area, namely, stability without subsequent input. According to R&D theory, we can let the representation be the state variable for the system whose dynamics are characterized in this manner.

Mathematically, these dynamics are easy to express with a differential equation: $\dot{x}(t) = \int u(t)\, dt$. In words, this system acts as a kind of integrator. In fact, neural systems with this kind of dynamics are often called "neural integrators" and are found in a number of brain areas, including brainstem, frontal lobes, hippocampus, and parietal areas. Neural integrators act like memories because when there is no input

[25] For a detailed description and review of these experiments and their results, see C. Colby and M. Goldberg, "Space and Attention in Parietal Cortex," *Annual Review of Neuroscience*, XXII (1999). 319–49, and R. Andersen, L. Snyder, D. Bradley, and J. Xing, "Multimodal Representation of Space in the Posterior Parietal Cortex and Its Use in Planning Movements," *Annual Review of Neuroscience*, XX (1997): 303–30.

(that is, $u(t)=0$), the change in the output over time is 0 (that is, $\dot{x} = \frac{dx}{dt} = 0$). Thus, such systems are stable with no subsequent inputs. Looking for the moment at Figure 1, we can see that the desired values of the **A** and **B** matrices will be 0 and 1 respectively in order to implement a system with these dynamics. Since we have a means of "translating" this canonical control system into one that respects neural dynamics, we can determine the values of **A**′ and **B**′ in Figure 2; they turn out to be 1 and τ (the time constant of the intrinsic neural dynamics), respectively. We can now set about building a model of this system at the level of single spiking neurons which gives rise to these dynamics—originally described at a higher level. In fact, the representation in LIP is far more complex than this, but the representational characterization of R&D theory is a general one, so such complexities are easily incorporated. As well, more complex dynamics are often necessary for describing neural systems, but again, the generality of R&D theory allows these to be incorporated using similar techniques.[27] So, while the neural integrator model is extremely simple, it shows how R&D theory provides a principled means of explaining a cognitive behavior (that is, memory) in a neurally plausible network.

II.4. Three principles. R&D theory is succinctly summarized by three principles:

(1) Neural representations are defined by the combination of nonlinear encoding (exemplified by neuron tuning curves) and weighted linear decoding.
(2) Transformations of neural representations are functions of variables that are represented by neural populations. Transformations are determined using an alternately weighted linear decoding.
(3) Neural dynamics are characterized by considering neural representations as control theoretic state variables. Thus, the dynamics of neurobiological systems can be analyzed using control theory.

To summarize these principles, Figure 3 shows a "generic neural subsystem." This figure synthesizes the previous characterizations of representation, computation, and dynamics, across multiple levels of description.

While recent, this approach has been successfully used to characterize a number of different systems including the vestibular system, the lamprey locomotive system, the eye stabilization system, working

[27] For various examples, see Eliasmith and Anderson, especially chapters 6 and 8.

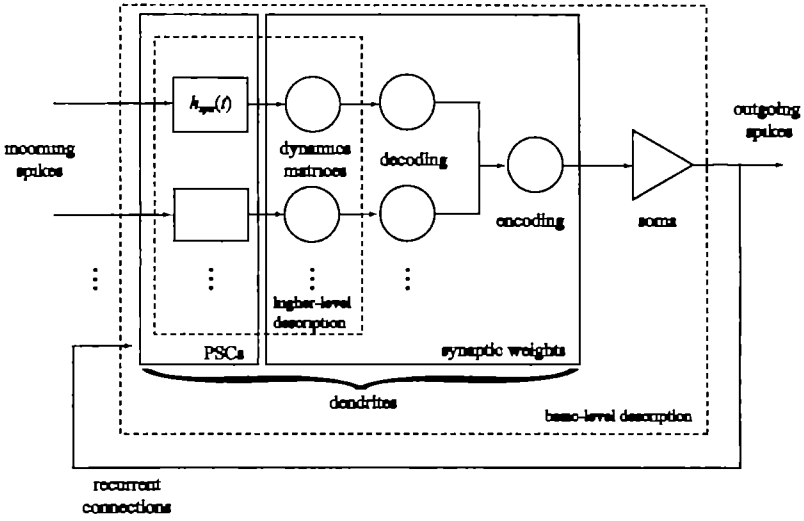


Figure 3: Generic neural subsystem. A synthesis of the preceding characterizations of representation (encoding/decoding), computation (biased decoding), and dynamics (captured by $h_{syn}(t)$ and the dynamics matrices). Dotted lines distinguish basic-level (that is, neural) and higher-level (that is, mathematical objects with units) descriptions.

memory,[28] and the limb control system.[29] As well, a rigorous formulation of each of these principles is available.[30]

III. COMPARISON TO CURRENT APPROACHES

Notice that R&D theory is not a description of cognitive systems as "being like" anything; cognitive systems are neurobiological systems that are best described by certain well-established quantitative tools for describing physical systems. That is, I am not proposing an analogy of "the mind as neural integrator" or even "the mind as a control system." While we may notice that certain control structures mimic some behaviors of neurobiological systems, we have to build the detailed neurobiological model and then determine if the mind really does work that way. In other words, R&D theory in no way suggests we should stop at the analogy. Rather, it gives us principled means of

[28] These examples can all be found in Eliasmith and Anderson.

[29] As described in Z. Nenadic, Anderson, and B. Ghosh, "Control of Arm Movement Using a Population of Neurons," *Mathematical and Computer Modelling*, xxxv (2002). 1261–69.

[30] This formulation can be found in Eliasmith and Anderson, pp. 230–31.

comparing a full-blown, neuron-level implementation with the actual neurobiological system. Connectionists, symbolicists, and dynamicists typically build models that begin and end with metaphors; in particular, they do not specify how their models relate to the mind as a physical system. But, the devil is in the details.

To see what is gained by moving away from the various metaphors that have dominated theorizing about the mind, let me briefly compare R&D theory to past approaches. In doing so, I describe several of what I take to be the central strengths and weaknesses of each approach. These lists are not intended to be exhaustive, but representative. So long as I am right about at least one of each, my claim that R&D theory should be preferred follows.

III.1. Symbolicism. The central problem for symbolicists, which I have already mentioned, is that time is largely ignored, or considered only after the fact.[51] Other typical concerns about symbolic models include their brittleness (that is, lack of ability to survive partial destruction of the system or its representations), high-level discreteness,[52] and unconvincing descriptions of low-level perceptual processes.[53] R&D theory suffers from none of these limitations due to its essential inclusion of time and its responsibility to the underlying neural architecture.

The central strengths of symbolicism are demonstrated by its many past successes (for example, ACT and SOAR). These are largely due to its espousal of cognitivism, that is, its willingness to peer inside the black box. Doing so has made representation an essential ingredient for providing good explanations of cognitive systems. As well, symbolicism is supported by a powerful and general theory of computation. Together, the commitment to representation and the quantitative theory of computation, make for a unified understanding of cognitive systems. R&D theory shares similar strengths. While computational theory is bolstered by including control and information theory, and the notion of representation is sharpened to relate directly to neural

[51] Perhaps the most concerted effort to include time in such models is in Allen Newell's *Unified Theories Of Cognition* (Cambridge: Harvard, 1990). This attempt, however, is both after-the-fact and largely inconsistent, as discussed in my "The Third Contender: A Critical Examination of the Dynamicist Theory of Cognition," *Philosophical Psychology*, IX, 4 (1996). 441–63.

[52] As discussed in E. Smith, "Concepts and Induction," in Michael Posner, ed., *Foundations of Cognitive Science* (Cambridge. MIT, 1989), pp. 501–26.

[53] This is made clear in the case of visual processes by P Churchland, V. Ramachandran, and T. Sejnowski, "A Critique of Pure Vision," in Christof Koch and Joel L Davis, eds., *Large-Scale Neuronal Theories of the Brain* (Cambridge: MIT, 1994), pp 23–60.

systems, the ability to provide unified representational explanations remains.

That being said, a typical concern of symbolicists regarding approaches that are concerned with neural implementations, is that the demonstrated symbol manipulating abilities of cognitive systems are lost in the concern for neural detail. In one sense, this issue is easily addressed in the context of R&D theory. This is because the numeric representations in R&D theory are just another kind of syntax. While it is not a typical syntax for the logic used to describe cognitive function by symbolicists, the syntax itself does not determine the kinds of functions computable with the system.[34] Given the discussion in section III.2, we know that this theory supports quite general computation, that is, linear and nonlinear functions. As a result, most, if not all, of the functions computed with standard symbolicist syntax can be computed with the numerical syntax adopted by R&D theory. More to the point, perhaps, past work using numerical distributed representations has shown that structure sensitive processing of the kind demanded by Fodor and Zenon Pylyshyn[35] can be achieved in such a system.[36] Furthermore, this kind of representational system has been used to model high-level cognitive functions, like analogical mapping.[37] As a result, structured symbol manipulation is not lost by adopting R&D theory. Rather, a precise neural description of such manipulation is gained.

III.2. Connectionism. Of past approaches, connectionism is probably the most similar to R&D theory. This raises the question: Is R&D theory merely glorified connectionism? A first response is to note that glorifying connectionism (that is, making it more neurally plausible) is no mean feat. The neural plausibility of many connectionist models leaves much to be desired. Localist models are generally not neurally plausible at all. But even distributed models seldom "look" much like real neurobiological systems. They include neurons with continuous, real-valued inputs and outputs, and often have purely linear or generic sigmoid response functions. Real neurobiological networks have highly

[34] This general point has been argued in J. Girard, "Proof-Nets: The Parallel Syntax for Proof-Theory," in Aldo Ursini and Paulo Agliano, eds., *Logic and Algebra* (New York: Marcel Dekker, 1996), pp. 97–124; and J. Girard, "Linear Logic," *Theoretical Computer Science*, L, 1 (1987): 1–102.

[35] "Connectionism and Cognitive Architecture: A Critical Analysis," *Cognition*, XXVIII (1988): 3–71.

[36] As demonstrated in T. Plate, "Distributed Representations and Nested Compositional Structure" (Ph.D. diss., University of Toronto, 1994).

[37] As in Eliasmith and Thagard, "Integrating Structure and Meaning: A Distributed Model of Analogical Mapping," *Cognitive Science*, XXV, 2 (2001): 245–86.

heterogeneous, nonlinear, spiking neurons. Connectionists themselves are seldom certain precisely what the relation is between their models and what goes on the brain.[38]

Of course, such connectionist models are far more neurally plausible than symbolicist ones. As a result, they are not brittle like symbolic systems, but rather degrade gracefully with damage. As well, they are supremely statistically sensitive and are thus ideal for describing many perceptual and cognitive processes that have eluded symbolicists. And finally, connectionist models do, on occasion, consider time to be central to neural processing.[39] Again, R&D theory shares each of these strengths and, in fact, improves on a number of them (for example, neural plausibility, and the integration of time).

But, more importantly, R&D theory also goes beyond connectionism. Connectionism has been predominantly a bottom-up approach to cognitive modeling. The basic method is straightforward: connect simple nodes together and train them to compute complex functions. While this approach can provide some useful insights (for example, determining what kinds of statistical structure can be detected in the training set), it is unlikely to lead to a useful model of a brain that consists of billions of neurons. Connecting ten billion nodes together and training them will probably not result in much. So, one of the main difficulties that connectionism suffers from is the lack of a principled method.

Progress in decomposing complex physical systems often necessitates an integration of bottom-up and top-down information.[40] So, in the case of neurobiology, it is essential to be able to test top-down hypotheses regarding brain function that are consistent with known lower-level facts. That is, we must be able to relate high-level characterizations of psychological processes (for example, "working memory") to more specific implementational claims (for example, that networks of certain kinds of neurons can realize such processes). Connectionists, unfortunately, have no principled method for incorporating top-down constraints on the design and analysis of their models. R&D theory, in contrast, explicitly combines both higher- and lower-level constraints on models.

[38] See the various discussions by the editors in McClelland and Rumelhart, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Volume 2.

[39] See the selection of the examples in Patricia S. Churchland and Terrence J. Sejnowski, *The Computational Brain* (Cambridge: MIT, 1992).

[40] As argued by Bechtel and Robert C. Richardson in their *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research* (Princeton: University Press, 1993).

The third principle of R&D theory captures this synthesis. It is with this principle that the analyses of representation, computation, and dynamics come together (see Figure 3). As an example, consider the recent proposal by Rajesh Rao and Dana Ballard[41] that the visual system acts like a dynamic, optimal linear estimator (that is, a linear control structure known as a Kalman filter).[42] Using R&D theory, we can build a large-scale, complex network to test this hypothesis. This is because the hypothesis is a precise high-level description, there is a significant amount of neural data available regarding the visual system, and principle three tells us how to combine these. Using the tools of connectionism, we simply cannot test this kind of high-level claim. It is not at all evident how we can train a network to realize an optimal estimator, or what an appropriate network architecture would be. So, R&D theory is able to test high-level hypotheses in ways not available to connectionists. This makes R&D theory much better able to bridge the gap between psychological and neural descriptions of behavior than connectionism.

Another way of making this point is to contrast the kind of characterization of dynamics principle three offers, with that typical of connectionism. While connectionists often consider the importance of time, and, in particular, have introduced and explored the relation between recurrent networks and dynamic computation, they do not have a systematic means of analyzing or constructing networks with these properties. Principle three, by adopting control theory, makes such analyses possible within R&D theory. That is, control theory has a large set of well-established quantitative tools for both analyzing and constructing control structures. And, because R&D theory provides a means of intertranslating standard and "neural" control structures, such tools can be used in a neurobiological context. This is extremely important for understanding the dynamic properties, and otherwise predicting the overall behavior of a network constructed using the R&D approach. In other words, R&D relates rather imprecise connectionist notions like 'recurrence' to a specific understanding of the dynamics of physical systems that is subject to well-known analytical tools. This makes it possible to design networks rigorously, with highly

[41] "Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects," *Nature Neuroscience*, II, 1 (1999): 79–87

[42] Another high-level hypothesis regarding the use of the Kalman filter has been made in the context of the construction and use of cognitive maps in hippocampus in K. Balakrishnan, O. Bousquet, and V. Honavar, "Spatial Learning and Localization in Animals: A Computational Model and Its Implications for Mobile Robots," *Adaptive Behavior*, VII, 2 (1999): 173–216.

complex dynamics (like the Kalman filter mentioned earlier), a task left mostly to chance with connectionism.

The previous discussion shows how principle three supports building networks that have the complex dynamics demanded by a higher-level hypothesis. In addition, principle three supports building networks that have the complex representations demanded by a higher-level hypothesis. For example, Eliasmith and Anderson[43] describe a model of working memory that accounts for representational phenomena not previously accounted for. Specifically, this model employed complex representations to demonstrate how working memory could be sensitive not only to spatial properties (that is, position) but to other properties concurrently (for example, shape). In addition, the model gives rise to both neural predictions (for example, connectivity and firing patterns), and psychological predictions (for example, kinds of error and conditions for error). This was possible only because R&D theory provides a means of determining the detailed connection weights given high-level descriptions of the system. Again, it is unclear how such a network could have been learned (that this is not an easy task is a good explanation for why it had not been previously done).

In both of these examples, R&D theory is distinguished from connectionism because it does not share the same heavy reliance on learning for model construction. Unfortunately, getting a model to learn what you want it to can be extremely challenging, even if you build in large amounts of innate information (and choosing what to build in tends to be something of an art). But, connectionists have little recourse to alternative methods of network construction, so the severe limitations and intrinsic problems with trying to learn complex networks are an inherent feature of connectionism. R&D theory, in contrast, allows for high-level characterizations of behavior to be imposed on the network that is constructed. As a result, connection weights can be analytically determined, not learned.

Nevertheless, R&D theory is also able to incorporate standard learning rules.[44] And, more than this, R&D theory can provide new insights regarding learning. This is because being able analytically to construct weights also provides some insight into methods for deconstructing weights. So, given a set of learned weights, the techniques of R&D theory can be used to suggest what function is being instantiated by

[43] "Beyond Bumps: Spiking Networks that Store Smooth N-Dimensional Functions," *Neurocomputing*, XXXVIII (2001): 581–86.

[44] As discussed in chapter 9 of Eliasmith and Anderson, *Neural Engineering: Computation, Representation and Dynamics in Neurobiological Systems*.

the network.[45] Often, when some input/output mapping has been learned by a connectionist network, it is very difficult to know exactly which function has been learned because the testable mappings will always be finite. Using R&D to determine which linear decoders combine to give a set of provided connection weights can be used to give an exhaustive characterization of what higher-level function is actually being computed by the network.

Such connection weight analyses are possible because R&D theory, unlike connectionism, explicitly distinguishes the encoding and decoding processes when defining representation and computation. While the relation between this distinction and the observable properties of neurobiological systems is subtle, as noted in section III.2, the theoretical payoff is considerable.

To summarize, R&D theory should be preferred to connectionism for two main reasons. First, R&D provides for a better understanding of neural connection weights, no matter how they are generated. There is much less mystery to a network's function if we have a good means of analyzing whatever it is that determines that function. While learning is powerful, and biologically important, it cannot be a replacement for understanding what, precisely, a network is doing. Second, R&D theory provides a principled means of relating neural and psychological data. This makes representationally and dynamically complex cognitive phenomena accessible to neural level modeling. Given a high-level description of the right kind, R&D theory can help us determine how that can be realized in a neural system. Connectionists, in contrast, do not have a principled means of relating these two domains. As a result, high-level hypotheses can be difficult to test in a connectionist framework.

So, unlike connectionism, R&D theory carefully relates neural and psychological characterizations of behavior to provide new insights into both. And, while it is possible that certain hybrid models (either symbolicist/connectionist hybrids, or localist/distributed hybrids) may make up for some of the limitations of each of the components of the hybrid alone, there is an important price being paid for that kind of improvement. Namely, it becomes unclear precisely what the cognitive theory on offer is supposed to be. R&D theory, in contrast, is highly unified and succinctly summarized by three simple, yet quantifiable, principles. To put it simply, Occam's razor cuts in favor of R&D theory. But, it should be reiterated that this unification buys

[45] For a simple example, see Eliasmith and Anderson, *Neural Engineering*, pp. 294–98.

significantly more than just a simpler theory. It provides a unique set of conceptual tools for relating, integrating, and analyzing neural and psychological accounts of cognitive behavior.

III.3. Dynamicism. Of the three approaches, dynamicism, by design, is the most radical departure from the "mind as computer" metaphor. In some ways, this explains both its strengths and its weaknesses. Having derided talk of representation and computation, dynamicists have put in their place talk of "lumped parameters," and "trajectories through state-space." Unfortunately, it is difficult to know how lumped parameters (for example, "motivation" and "preference")[46] relate to the system that they are supposed to help describe. While we can measure the arm angle of the Watt Governor, it is not at all clear how we can measure the "motivation" of a complex neurobiological system. But this kind of measurement is demanded by dynamicist models. As well, some dynamicists insist that cognitive models must be low-dimensional, in order to distinguish their models from those of connectionists.[47] But insistence on low-dimensionality greatly reduces the flexibility of the models, and does not seem to be a principled constraint.[48] Finally, because the Watt Governor, a standard example of a classical control system, has been chosen as a central exemplar of the dynamicists approach, the well-known limitations of classical control theory are likely to plague dynamicism. Clearly, these limitations are not ones shared by R&D theory.

What the replacement of the "mind as computer" metaphor by the "mind as Watt Governor" metaphor gained for dynamicists was an appreciation of the importance of time for describing the behavior of cognitive systems. No other approach so relentlessly and convincingly presented arguments to the effect that cognitive behaviors are essentially temporal.[49] If, for example, a system cannot make a decision before all of the options have (or the system has) expired, there is little sense to be made of the claim that such a system is cognitive.

[46] These are two of the lumped parameters included in the model of feeding described in J. Busemeyer and J. Townsend, "Decision Field Theory. A Dynamic-Cognitive Approach to Decision Making in an Uncertain Environment," *Psychological Review*, c, 3 (1993): 432–59.

[47] van Gelder, "What Might Cognition Be, If Not Computation?"

[48] These points are discussed in detail in my "Commentary: Dynamical Models and van Gelder's Dynamicism: Two Different Things," *Behavioral and Brain Sciences*, xxi, 5 (1998): 615–65, and my "The Third Contender: A Critical Examination of the Dynamicist Theory of Cognition."

[49] Such arguments are prominent in both van Gelder's "What Might Cognition Be, If Not Computation?" and his "The Dynamical Hypothesis In Cognitive Science," *Behavioral and Brain Sciences*, xxi, 5 (1998) 615–65, and the various contributions to Port and van Gelder.

Furthermore, there is evidence that perception and action, two clearly temporal behaviors, provide the foundation for much of our "more cognitive" behavior.[50] While dynamicists have done a good job of making this kind of argument, the consequences of such arguments need not include the rejection of representation and computation that dynamicists espouse. R&D theory, which essentially includes precisely these kinds of dynamics, shows how representation, computation, and dynamics can be integrated in order to tell a unified story about how the mind works.

III.4. Discussion. So, in short, R&D theory adopts and improves upon the dynamics of dynamicism, the neural plausibility of connectionism, and the representational commitments of symbolicism. As such, it is a promising synthesis and extension of past approaches to understanding cognitive systems, because it includes the essential ingredients. Of course, it is not clear whether R&D theory combines those ingredients in the right way. Because it is a recent proposal for explaining cognitive systems, its current successes are few. While it has been used to effectively model perceptual (the vestibular system), motor (eye control), and cognitive (working memory) processes, these particular examples of perceptual, motor, and cognitive behavior are relatively simple. So, while the resources for constructing neurally plausible models of phenomena that demand complex dynamics over complex representations are available, it remains to be clearly demonstrated that such complexity can be incorporated into R&D theoretic models.

As well, R&D theory does not, in itself, satisfactorily answer questions regarding the semantics of representational states. As Fred Dretske[51] has noted, coding theory does not solve the problem of representational semantics. Thus, R&D theory needs to be supplemented with a theory of meaning, as mentioned in section III.1. In fact, I think R&D theory suggests a novel theory of meaning that avoids the problems of past theories.[52] Nevertheless, this remains to be clearly demonstrated.

Even in this nascent form, however, R&D theory has some important theoretical implications for work in philosophy of mind and cognitive science. For example, functionalism regarding the identity of mental

[50] This view, associated variously with terms "embodied," "embedded," or "dynamicist" has been expressed in, for example, Francisco J. Varela, Evan Thompson, and Eleanor Rosch, *The Embodied Mind: Cognitive Science and Human Experience* (Cambridge: MIT, 1991), and more recently in Ballard, M. Hayhoe, P. Pook, and R. Rao, "Deictic Codes for the Embodiment of Cognition," *Behavioral and Brain Sciences*, in press

[51] *Knowledge and the Flow of Information* (Cambridge: MIT, 1981).

[52] For an attempt at articulating such a theory, see my "How Neurons Mean: A Neurocomputational Theory of Representational Content"

states may need to be reconceived. If, as R&D theory entails, the function of a mental state must be defined in terms of its time course, and not just its inputs and outputs, it is unlikely that functional isomorphism of the kind that Hilary Putnam[53] envisioned will be sufficient for settling the identity of mental states. If the dynamics of some aspects of mental life are central to their nature, then an atemporal functionalism is not warranted. Standard functionalism in philosophy of mind is clearly atemporal. And, I take it, some (if not many) aspects of mental life have their character in virtue of their dynamics (for example, shooting pains, relaxed conversations, and recognizing friends). So, a "temporal" functionalism is necessary for properly characterizing minds. In other words, input, outputs, and their time course must all be specified to identify a mental state. While some mental functions may not be especially tied to dynamics (for example, addition), others will be (for example, catching a ball). Specifying ranges of dynamics that result in the successful realization of that function will allow us to determine if some mind or other can really be in a given mental state.

These considerations have further consequences for the role of the Turing machine in cognitive science.[54] While cognitive functions will still be Turing computable, they will not be realizable by every universal machine. This is because computing over time (that is, with time as a variable in the function being computed) is different from computing in time (that is, arriving at the result in a certain time frame). When this difference is acknowledged, it becomes clear that Turing machines as originally conceived (that is, under the assumption of infinite time) are relevant theoretically, but much less so practically (that is, for understanding and identifying real minds). I take it that more argument is needed to establish such conclusions, but that, at the very least, adopting R&D theory shows how such positions are plausible.

IV. CONCLUSION

Perhaps, then, R&D theory or something like it can help rid us of the constraints of metaphorical thinking. Such an approach holds promise for preserving many of the strengths, and avoiding many of the weaknesses, of past approaches to understanding the mind. But,

[53] "Philosophy and Our Mental Life," in his *Mind, Language, and Reality: Philosophical Papers* (New York: Cambridge, 1975), pp. 291–303

[54] These consequences are more fully explored in my "The Myth of the Turing Machine: The Failings of Functionalism and Related Theses," *Journal of Experimental and Theoretical Artificial Intelligence*, XIV (2002): 1–8

THE END OF COUNTERPART THEORY*

Counterpart theory says roughly that, for any object O and any property F, O is possibly F if and only if O has a counterpart that is F. Moreover, O is essentially F if and only if all of O's counterparts are F.[1] According to David Lewis,[2] the theory's leading advocate, our counterparts are typically a lot like us. Lewis holds that I am possibly forty feet tall if and only if there is someone in a universe spatiotemporally isolated from ours—one of Lewis's "possible worlds"—who, though otherwise appropriately like me, is forty feet tall.

Many find counterpart theory attractive, but most reject Lewis's modal realism.[3] So most deny that we have flesh-and-blood counterparts in unreachable but humanly inhabited universes. They insist, instead, that our counterparts are somehow "abstract." It is that sort of counterpart theory—the sort endorsed by virtually every counterpart theorist except for Lewis himself—that I shall argue is untenable. (Indeed, as we shall see, there are good reasons to reject *any* reduction of modal properties to abstract worlds, counterpart-theoretic or otherwise.) Because I do not believe Lewis's ontology, I think his version of counterpart theory is also mistaken. And so—I conclude—we should reject every sort of counterpart theory.

I. COUNTERPART THEORY AND THE ANALYSIS OF MODAL PROPERTIES

The counterpart theorist says that, for any object O and any property F, O is possibly F if and only if O has a counterpart that is F. But she

* Thanks to Bob Adams, Jim Cargile, Dave Chalmers, Jan Cover, Michael Della Rocca, John Devlin, John Divers, Geoff Goddu, Mark Heller, Harold Langsam, Gene Mills, Mark Murphy, Al Plantinga, Gideon Rosen, Dean Zimmerman and, especially, Mike Bergmann, Mike Rea, and Ted Sider. Versions of this paper were presented at Virginia Commonwealth University, Western Washington University, the University of Virginia, and Metaphysical Mayhem VII. This paper was supported, in part, by a summer grant from the University of Virginia.

[1] This characterization is rough, leaving out, among other things, the sortal-relative nature of counterpart relations.

[2] See *On the Plurality of Worlds* (New York: Blackwell, 1986).

[3] Advocates of counterpart theory include Graeme Forbes ("Two Solutions to Chisholm's Paradox," *Philosophical Studies*, XLVI (1984): 171–87), Alan Gibbard ("Contingent Identity," *Journal of Philosophical Logic*, IV (1975): 187–221); Allen Hazen ("Counterpart-Theoretic Semantics for Modal Logic," this JOURNAL, LXXVI, 6 (June 1979): 319–38); Mark Heller ("Property Counterparts in Ersatz Worlds," this JOURNAL, XCV, 6 (June 1998): 293–316); Theodore Sider (*Four-Dimensionalism: An Ontology of Persistence and Time* (New York: Oxford, 2001), pp 111–13); and Robert Stalnaker

0022-362X/03/0010/521–49 © 2003 The Journal of Philosophy, Inc.

does not believe that the having of counterparts' going hand in hand with the having of modal properties is some sort of cosmic modal coincidence. Her idea is not that, as a synthetic necessity, counterpart relations merely but unerringly "track" modal properties. Rather, the counterpart theorist thinks that *O*'s having a counterpart that is *F* is *what it is* for *O* to be possibly *F*. Standing in the relevant counterpart relation is *what it is* to have the "corresponding" modal property. Thus, counterpart theory is about *analyzing modal properties*. (Or it is about analyzing possibilities. Although I shall usually describe counterpart theory as analyzing *O*'s property of *being possibly F*, we could just as well describe it as analyzing *the possibility of O's being F*.)

We can further illustrate this point—that counterpart theory is in the business of providing counterpart-theoretic analyses of modal properties—by considering the "classic" objection to counterpart theory. That objection charges counterpart theory with changing the subject. When I ask whether I might have been happier, so the objection goes, I am asking whether *I*—this very person—might have been happier. It is simply not to answer my question to say that an otherworldly someone else is happier, even if he is very much like me, even if we call him "Merricks's counterpart."

This sort of objection was raised early on by both Alvin Plantinga[4] and Saul Kripke.[5] Plantinga says:

> According to [counterpart theory] there is a world W...that includes the existence of an unwise counterpart of Socrates...and how is that even relevant to the claim that Socrates himself—the Socrates of [our world]—could have been unwise? There could have been a foolish person a lot like Socrates; how does this fact show that *Socrates* could have been unwise (*op. cit.*, p. 116)?

And consider the following from Kripke:

> The counterpart of something in another possible world is *never* identical with the thing itself. Thus if we say 'Humphrey might have won the election (if only he had done such-and-such)', we are not talking about something that might have happened to *Humphrey* but to someone else, a "counterpart." Probably, however, Humphrey could not care less whether someone *else*, no matter how much resembling him, would have been victorious in another possible world. Thus, Lewis's view seems to me even more bizarre than the usual notions of transworld identification that it replaces (*op. cit.*, p. 45, note 13).

("Counterparts and Identity," *Midwest Studies in Philosophy*, XI (1986): 121–40)

[4] *The Nature of Necessity* (New York: Oxford, 1974).

[5] *Naming and Necessity* (Cambridge: Harvard, 1980).

The counterpart theorist's rejoinder is that the unwise counterpart of Socrates is relevant to Socrates's being possibly unwise because Socrates's having such a counterpart is *one and the same thing* as his being possibly unwise. Similarly, she insists, the very thing Humphrey cares about—possibly winning the election—just is, is analyzed as, his having a counterpart who wins.[6]

Now Humphrey might *think* he does not care about having a victorious counterpart. But Humphrey's thinking this, the counterpart theorist maintains, does not imply that counterpart theory is false. It implies, instead, that Humphrey is mistaken about how to analyze modal properties. Humphrey's situation, the counterpart theorist should claim, is like that of the parched man who says, "I could not care less about H_2O; I want water!"

The counterpart theorist's claim here is, of course, controversial. Obviously, those who hold that the classic objection itself undermines counterpart-theoretic analyses of modal properties will look askance on any response that trades on such analyses. Perhaps we are at an impasse, since, just as obviously, the counterpart theorist will insist that such analyses are correct. But all that matters for our purposes is that the counterpart theorist will defend such analyses.

II. NO "GENERIC" COUNTERPART-THEORETIC ANALYSES

The counterpart theorist must say that there are counterpart-theoretic analyses of modal properties. But any such analysis, if it is to have a hope of being successful, is committed to a substantive account of the nature of counterparts and possible worlds. To see why I say this, consider—for reductio—a purported analysis free of any such commitments. Consider this purely "generic" counterpart-theoretic analysis:

(1) *O* is possibly forty feet tall *is analyzed as* *O* has a counterpart that is forty feet tall.[7]

In these debates, 'counterpart' is a technical term, its meaning stipulated by the counterpart theorist. And *something* must be stipulated; if nothing at all were stipulated, (1) would be meaningless rather than merely generic. Moreover, if (1) is to be generic, the stipulated meaning cannot favor one account of counterparts over another.

[6] Cf. Hazen, p. 322, Lewis, p 196, and Sider, p. 195.

[7] One might recast (1) as an account of *O*'s being *merely* possibly forty feet tall This change has some benefits. For example, it allows *O*'s being actually forty feet tall, all by itself, to make *O* possibly forty feet tall, no counterpart needed. This is not relevant to my arguments, so I shall ignore it in what follows.

The only meaning that seems to fit the bill—the only thing about counterparts upon which all accounts agree—is the following. An object *O*'s counterpart is that which, by having a nonmodal property, thereby gives the "modal version" of that property to *O*.[8]

So (1), if truly generic, is really a condensed version of the following:

(2) *O* is possibly forty feet tall *is analyzed as* there is a forty-foot-tall something other than *O* whose being forty feet tall is *what it is* for *O* to be possibly forty feet tall.

(In order to accommodate abstract counterpart theory, (2) itself should be read as a gloss of the more careful:

(2*) *O* is possibly forty feet tall *is analyzed as* there is something other than *O* that stands in some relation to the property of *being forty feet tall* and whose being thus related to that property is *what it is* for *O* to be possibly forty feet tall.

If *O*'s counterpart is an abstract object it will have no height whatsoever. So it discharges its duty by standing in some relation other than exemplification to *being forty feet tall*. But for ease of exposition and only when it will cause no confusion, I shall continue to speak as if an object's alleged counterparts have the properties that it possibly has.)

Even by the counterpart theorist's own lights, (2) is not a good analysis of *O*'s being possibly forty feet tall. For *O*'s being possibly forty feet tall is invoked in the analysans. Thus, (2) is unacceptably circular (likewise with (2*)). Given only generic counterpart theory, (1) fares no better. For, as noted, (1)'s use of 'counterpart' makes it merely a condensed statement of (2).

(1) is done in by its use of 'counterpart'. But we cannot repair (1) by simply dropping that word. For consider:

(3) *O* is possibly forty feet tall *is analyzed as* there is something other than *O* that is forty feet tall.

Even if counterpart theory is correct, (3) is not. Not just any old forty-foot-tall object will do the trick. (The result of making the corresponding change to (2*) is even worse:

(3*) *O* is possibly forty feet tall *is analyzed as* there is something other than *O* that stands in some relation to the property of *being forty feet tall*.)

The analysantes of (1) and (2) say too much; that of (3) says too

[8] Cf. Sider, "Review of Michael Jubien's *Ontology, Modality, and the Fallacy of Reference*," *Noûs*, XXXIII (1999): 284–94, here p. 290.

little. To get a sense of how to avoid both extremes, contrast (1) through (3) with an analysis of the sort Lewis might give:

(4) *O* is possibly forty feet tall *is analyzed as* there is something other than *O*, in a universe spatiotemporally isolated from ours, that is appropriately similar to *O* but that is forty feet tall.

This analysis will prompt further questions, such as: What does 'appropriately similar' amount to? But it is not circular, giving it a leg up on (1) and (2). Moreover, (4) says enough about the object invoked in the analysans to separate it from the obviously inadequate (3). Thus (4), unlike (1) through (3), is at least a legitimate contender.

The counterpart theorist need not agree with Lewis on the nature of counterparts; she need not endorse (4). But if the counterpart theorist hopes to analyze modal properties successfully, she must avoid the errors of (1) through (3) and emulate the virtues of (4). And to do that, she must tell us something about what alleged counterparts are like. There can be no generic counterpart-theoretic analyses of modal properties.

Those who write in any detail about counterpart theory typically say something about the nature of counterparts. Often, they emphasize that our counterparts need not be concrete human beings located in physical worlds; they emphasize that counterpart theory as such does not imply Lewis's extravagant ontology.[9] But this had better not be the end of their story. For no counterpart-theoretic analyses of modal properties are endorsed by saying only that counterpart theory is true but Lewis got it wrong. After all, the implicit circularity afflicting (1) also afflicts:

(1*) *O* is possibly forty feet tall *is analyzed as O* has an *abstract* counterpart that is forty feet tall.

I have considered various attempts at a generic analysis that begin with the idea that our counterparts have the properties that we possibly have. We have seen the problems with that approach. But before leaving attempted generic analyses behind, there is one other approach worth noting. An aspiring generic analyst might offer:

(5) *O* is possibly forty feet tall *is analyzed as O* is represented as being forty feet tall in an abstract possible world.

(5) does not invoke the term 'counterpart' in its analysans. Thus (5)

[9] See, for example, Hazen, pp 319–20; and Sider, "Review of Michael Jubien's *Ontology, Modality, and the Fallacy of Reference*," p. 290.

gains superiority over (1) and (1*). And it achieves generic status by failing to allude to counterparts by way of some other, more controversial, description. But these gains and achievements amount to leaving counterparts out altogether. And so (5) fails to deliver a bona fide counterpart-theoretic analysis, generic or otherwise.

In other words, (5) illegitimately presupposes that every analysis of a modal property in terms of what a possible world represents is counterpart-theoretic. (5) can be modified so that it rules out non-counterpart-theoretic approaches. But doing so involves specifying, in particular, *how* a world represents *O* as forty feet tall. As we shall see in what follows, this takes us down one or another controversial path. Once we start down one or another of these paths, we part ways with one or another counterpart theorist. Thus, our account ceases to be generic. (Saying a world represents *O* by having a *counterpart* of *O* would be suitably generic; but of course it reintroduces the circularity that undermined (1) and (1*).)

In the sections following, I shall attack nongeneric counterpart-theoretic analyses of modal properties. A defender of counterpart theory might be tempted to reply that objections based on this or that account of the nature of counterparts are not objections to counterpart theory as such. But, as we have seen, counterpart theory as such implies that there are counterpart-theoretic analyses of modal properties. Abstract counterpart theory as such is committed to analyses in terms of some detailed account or other of the nature of abstract counterparts and abstract worlds, analyses that take us beyond suggestions like (1) and (1*). If all those analyses fail, then so does abstract counterpart theory itself.

My objections to generic counterpart-theoretic analyses are part of my overall argument against counterpart theory. Their primary point is to prepare the way for that argument, to get us to see that the counterpart theorist as such is committed to there being some more detailed analysis or other of modal properties, something more detailed than the sketchy answers considered above. But my overall argument against counterpart theory aside, the objections above—all by themselves—should trouble most counterpart theorists. For most counterpart theorists have little more to say about the nature of counterparts than that they are "abstract." Thus the objections above show that most counterpart theorists have no counterpart-theoretic analyses of modal properties. And to endorse counterpart theory—which just is the theory that some such analyses are true—without understanding and endorsing any such analysis is to buy a pig in a poke, to leap without looking, to endorse a theory without knowing its content.

III. HELLER'S ABSTRACT COUNTERPART THEORY

Most abstract counterpart theorists say little about the nature of counterparts, but not all are so reticent. Mark Heller presents a detailed account of the nature of abstract worlds and counterparts (*op. cit.*). (Indeed, among defenders of abstract counterpart theory, Heller may be the most explicit about the details.) I shall raise five objections to Heller's counterpart-theoretic analyses of modal properties.

I shall develop these objections in some detail, in part because my target is not Heller alone. For after discussing Heller, I shall show how basically those same objections undermine almost every attempt at analyzing modal properties in terms of abstract counterparts. That is, I shall argue that they directly undermine all such attempts except those that rely upon *sui generis* propositions (or states of affairs, and so forth) (section IV). I then show why no abstract counterpart theorist can rely on such propositions (or states of affairs, and so forth) (section VI).

Heller begins to explain his view by relating an account of worlds sketched by W.V. Quine.[10] According to that account, a world is a set of ordered pairs. The first member of each ordered pair is an ordered quadruple representing a particular point in the "space-time manifold." The second is either a 1—representing that point as filled—or a 0—representing that point as empty. In order to allow for properties not supervenient on distributions of *being filled* or *being empty*, Heller modifies the Quinean account. Heller has the second member of each ordered pair be a set that represents a property; a different set represents each property that can be exemplified at a point in space time.[11]

So given Heller's theory, the following is a possible world:

SET *S*: {<<13, 69, 1560, 2>, {3, 48}>, <<589, 4, 37, 38>, {7, 19}>, <<77, 82, 99999, 311>, {3}>...}

Heller's worlds "represent" things as being a certain way (*op. cit.*, p. 307). Moreover, Heller explicitly endorses analyzing modal claims in terms of what his possible worlds represent (*op. cit.*, p. 315). For example, I am possibly forty feet tall is analyzed by Heller as my being

[10] Quine sketches this account in "Propositional Objects," in *Ontological Relativity and Other Essays* (New York: Columbia, 1969), pp. 139–60

[11] Heller says that these modifications do not take us all the way to a fully worked out account of worlds—see p. 316 For example, a full account would have a way of representing abstract objects, which do not exemplify properties at points in space time.

represented as being forty feet tall in some of his possible worlds. Thus we find in Heller an attempt at giving analyses of modal properties.

Those analyses are *counterpart-theoretic*. A world represents me as being forty feet tall in virtue of its having among its members a series of quadruples which thus represent me; that series is my "counterpart" in that world. (Moreover, its counterparts need not be my counterparts; thus we get the intransitivity of "transworld identity" that is a hallmark of counterpart theory.) Some may find this a strained notion of 'counterpart'. But, of course, any *abstract* counterpart theorist must invoke this or some equally "strained" notion.[12] And this is no problem. As already noted, in these debates 'counterpart' is a title of art; whatever plays the appropriate role in the analysis merits that title.

Heller's worlds are sets. We might wonder how a set like the one gestured at above manages to represent a possibility. Heller's answer is that sets represent only because we use them to do so. Moreover, which sets represent which possibilities depends on how the sets are interpreted. Indeed, Heller goes so far as to say, "if no one had ever interpreted sets as possible worlds, then sets would not have represented alternative possibilities" (*op. cit.*, p. 315).

Heller *should* say these things. Reconsider the view Quine originally sketched. There it is a matter of decision—not prior objective fact—that 1 represents a point as filled and 0 a point as empty. The number 1 could just as easily have represented being empty, 0 being filled. Likewise, given Heller's account of worlds, it is a matter of interpretation whether {3, 48} represents (for example) negative charge rather than some other property. And whether a series of quadruples (in a given world) is my counterpart representing me as tall, rather than yours representing you as short, is likewise a matter of interpretation.

Objection 1: Heller claims that being possibly forty feet tall is analyzed in terms of set-theoretic representations. And he claims that which sets represent which possibilities depends on our interpretations. Together, these claims imply an unacceptable conclusion. They imply that the nature of my being possibly forty feet tall is, among other things, a function of how we interpret sets. After all, given one interpretation of sets, my being possibly forty feet tall will be analyzed as (my existing and) the existence of set *S* above. Given another interpretation, my being possibly forty feet tall will be analyzed as the existence of a distinct set *T*, differing from *S* by (for example) replacing every occurrence of 3 with a 7, and every occurrence of 7 with a 3.

The point here, obviously, has nothing to do with my being possibly

[12] Cf. Lewis, p. 149.

forty feet tall in particular. If Heller is right, the nature of every modal property depends on something that we do. So I conclude that Heller is not right. For it is surely false that, for example, what it is for the sun to be possibly just a hair smaller is, even in part, a result of anything we do. This is my first objection to Heller.

Objection 2: Set aside my first objection. Grant Heller that, as a result of our interpretations, an object's having a modal property is analyzed in terms of the existence of such and such a set-theoretic entity. This implies—since our interpretations could have differed—that the analyses that hold as a result of those interpretations could have differed. And this in turn implies that the analysis of each modal property is itself a contingent matter. But the analysis of any property is a matter of necessity. That Heller's view implies otherwise is my second objection.

This objection is not troubled by the obvious fact that what we mean by the English expression 'is possibly forty feet tall' is contingent. There is no question that that expression could have meant that a certain set S exists, or instead that set T exists, or instead that chickens exist. Nevertheless, it is in fact true that 'is possibly forty feet tall' means exemplifies the property of *being possibly forty feet tall*. The counterpart theorist's task is to analyze properties like *being possibly forty feet tall*. And, to repeat, the true analysis is a matter of necessity.

Objection 3: My third objection begins by considering an objection Heller addresses in his article:

> In a world with no interpreters, the sets that actually are possible worlds go uninterpreted...since no worlds exist in worlds without interpreters, nothing is possible (or necessary) in those worlds. So if there had been no interpreters, there would have been no modal truths (*op. cit.*, p 314).

Heller takes this objection seriously. (This alone shows that Heller is promoting an analysis of modality; for this objection would be a howler if interpreted sets were supposed to be anything else.) His reply is that, once interpreted, sets are indeed possible worlds; and so, once interpreted, we can say truly that wherever those sets exist—even in worlds without interpreters—there are alternate possibilities (*op. cit.*, pp. 314–15). Suppose we grant Heller his reply. Suppose that once the interpreting is actually done, we have all the modality we want or need, even in worlds without interpretation.

I think that the interpreting has not all been done; and so I think that—given Heller's view—we lack the modality we need. To begin to see why I say this, note that it was not true a thousand years ago that humans interpreted sets to represent the possibility of molecules' being composed of atoms. For humans then were not able even to

entertain this possibility. But we are, surely, in a situation akin to that of our forbears. For there are, surely, possibilities we cannot entertain. And so, as a result, we have not interpreted sets to represent those possibilities. Similarly, we have not interpreted sets to represent every possibility involving possible properties and individuals that do not and did not and will not actually exist (the "alien" properties and individuals). So we have not—and will not and cannot—interpret sets to represent *all* the possibilities that we know there are.[13] So Heller's view—which implies that the extant possibilities cannot outstrip the interpreted sets—is wrong. This is my third objection to Heller.[14]

As an aside, I note that the above three objections threaten more than just abstract counterpart theory. They are objections to any view that endorses both of the following claims. First, *de re* (or even *de dicto*) modality is reduced to what abstract structures represent. Second, abstract structures represent in virtue of what we do. Even a noncounterpart theorist who wants to reduce modal properties to set-theoretic constructions is threatened by the above objections. Relatedly, the above objections—along with the one to follow—show that propositions are not sets, at least not sets relevantly like those Heller invokes. (For example, propositions stand in certain interrelations, such as entailment, essentially. Suppose set S is a proposition that entails set T because of what each set represents. And suppose each set represents what it does in virtue of what we contingently do. Then S entails T contingently. Absurd.)

Objection 4: The above objections turn on the role Heller places on interpretation. Some might object that this picks on the worst part of his view. They might say a better way to take Heller—or perhaps a better version of a Heller-like counterpart theory—allows the worlds to represent "objectively" rather than as a result of our interpreting. Or, alternatively, they might say that we should consider a Heller-style approach that leaves representation out of the picture altogether.

Let us begin with this second suggestion. Defenders of this suggestion might say, for example, that my being possibly forty feet tall is

[13] Cf. Lewis, p. 156.

[14] Consider the view that O's being possibly F is analyzed as *there being some set or other that is interpreted as representing O as F*. (That set will have to be "maximal" and "possible.") According to this view—unlike Heller's—our interpretations do not "select" the analysis of any property. Rather, our interpreting is part of the analysis itself, an analysis that is constant no matter how we interpret. This sort of approach is not subject to my second objection. It is subject to a cousin of my first objection. But most importantly, it is devastated by the third. For by building interpreting into the analysans of each modal property, it implies that there are no modal properties in worlds without interpreters.

analyzed as the existence of set *S*. But, they add, this has nothing to do with what *S* represents. They might even add that just as water's being H_2O has nothing to do with what H_2O represents, so my being possibly forty feet tall's being set *S* has nothing to do with what *S* represents. The identity or analysis or reduction in one case, just as in the other, is simply a straightforward metaphysical fact.

Taxation without representation is bad. Reduction without representation is worse, at least when it is the reduction of *de re* modal properties to abstract structures. For such alleged reductions are wholly unmotivated. It is *S*'s supposedly representing me as forty feet tall that motivates an account of my being possibly forty feet tall in terms of *S*. (Similarly, *S* qualifies as a possible world only because of what it represents, only because it is a maximal and consistent representation.) Of course, one can simply announce that my being possibly forty feet tall just is the existence of *S* apart from what—if anything—*S* represents. But this is like the bald assertion that my being possibly forty feet tall just is my being on the same planet as a yawning cat, or (instead) my being less than ten light years from the moon, or my being such that three is prime.

Even if we want to adjust Heller's theory so that it avoids the first three objections, we should not abandon representation. The other option, then, is to insist that the sets in question do represent, but not in virtue of anything we do. I am not sure how sets (of ordered pairs of ordered quadruples and sets of numbers) are supposed to represent thus "objectively." It is odd to think that a set of ordered pairs represents me as being forty feet tall and the sun as being a hair smaller and so on in and of itself, intrinsically and objectively, not on the basis of any sort of interpretation. It is strange to say that sets could—in and of themselves—pull this off. The suggestion seems to be that these sets are a language, but not by convention.

The oddness of such a view aside, we might wonder *how* these sets represent. One reply is that they just do, end of story. Aficionados of ersatz possibilia will see that this reply is like what Lewis calls "magical ersatzism" (*op. cit.*, pp. 174ff.). Presumably, if counterpart theory is built on a foundation of primitive or brute representation, it will be built on something like *sui generis* propositions (or states of affairs, and so forth) rather than on sets like *S*. So "magical" representation does not so much suggest a way of modifying Heller as it does an entirely different approach. And I shall address that approach below (section VI). (But, for the record, what I say about *sui generis* propositions easily applies to the claim that Heller's sets magically represent.)

Perhaps the defender of a Heller-style theory will insist that sets like *S* represent objectively, but deny that they do so in the primitive

way credited to *sui generis* propositions. The only likely suggestion here is that sets like Heller's represent various possibilities in virtue of their "structure." And so, along these lines, Theodore Sider says: "Modal concepts lay down a structural requirement: our talk of possible worlds and the rest is about any structure that is suited to play the relevant roles."[15] So let us consider whether we can get objective representation from sets like *S* via their "structure."[16]

One objection to this idea begins by considering Paul Benacerraf's famous argument[17] against identifying numbers with sets. A set's being (for example) the number three would be the result of some feature of that set. But there is no relevant feature that one of the sets competing to be three has and all the others lack, no difference of the sort necessary to make one set (but none of its competitors) the number three. So either none of the competitors is three, or more than one of them is. But it is false that more than one of them is. (By the transitivity of identity, if any set is (identical with) the number three, no set not identical with it is also (identical with) the number three.) So no set is the number three. Parts of Benacerraf's paper are quite controversial; but its least controversial conclusion is that there is no set *S* such that it is the number three.[18]

A similar sort of argument has been made against set-theoretic reductions of ordered pairs. Peter Forrest considers Kuratowski's construction

> ...whereby <a,b> is identified with {a,{a,b}}. But, as is well known, Kuratowski's identification suffers from a grave defect if it is treated as anything more than a model-theoretic device. For it is a convention, not a discovery, that <a,b> is to be identified with {a,{a,b}} rather than, say, {{a,ø},{a,b}}, and serious ontology is not done by convention.[19]

[15] "The Ersatz Pluriverse," this JOURNAL, XCIX, 6 (June 2002): 279–315, see p. 310.

[16] This view—the view that a world represents by its "structure"—resembles "pictorial" ersatzism, according to which a world represents "by isomorphism" (Lewis, pp. 166–67).

[17] "What Numbers Could Not Be," *Philosophical Review*, LXXIV (1965). 47–73.

[18] Benacerraf has many critics, including Crispin Wright (*Frege's Conception of Numbers as Objects* (Aberdeen: University Press, 1983)); Jerrold Katz ("Skepticism about Numbers and Indeterminacy Arguments," in Adam Morton and Stephen Stich, eds., *Benacerraf and His Critics* (Cambridge: Blackwell, 1996), pp. 119–39); Penelope Maddy ("Sets and Numbers," *Noûs*, XV (1981): 495–511); and Linda Wetzel ("That Numbers Could Be Objects," *Philosophical Studies*, LVI (1989): 273–92). Yet none dispute the points just noted. (A supervaluationist might say that 'some set is the number three' is true, claiming it is true on each precisification of 'three', but she will not say that there is a set *S* that is the number three, that is, she will not say that there is a set *S* such that '*S* is the number three' is true on all precisifications of 'three')

[19] "Neither Magic Nor Mereology: A Reply to Lewis," *Australasian Journal of Philosophy*, LXIV (1986): 89–91, here pp. 90–91.

The same point is made by D.M. Armstrong, who concludes of identifying ordered pairs with unordered sets of sets: "...as a piece of serious metaphysics, this seems quite unacceptable."[20]

With all this in mind, turn to the Heller-inspired theory according to which sets represent possibilities "objectively," by way of their structure. It cannot be that my being possibly forty feet tall *just is* or *is analyzed as* my being thus represented by set *S* and also that it is false—because it is analyzed in some other way—that my being possibly forty feet tall *just is* or *is analyzed as* my being thus represented. So, given the view here about the analysis of modal properties, there is a privileged set-theoretic way to represent my being possibly forty feet tall. But—certainly—there is not.

The objection here is not that this view implies that more than one series of quadruples will qualify as my forty-foot-tall counterpart. This implication is unproblematic; counterpart theory happily insists that I have many counterparts in multiple worlds, many of whom make me possibly forty feet tall. Rather, the objection here is that this approach absurdly implies that no single analysis of a modal property is better than some incompatible analyses.

To see this point more clearly, return to the idea that a world is a set of ordered pairs, the first member of which is an ordered quadruple locating a point, the second of which represents that point as filled or as empty. One way of doing things has 1 representing a point as filled and 0 as empty; another approach has it the other way around. Obviously, these approaches are inconsistent with each other: one approach uses 1 to represent being empty, the other uses it to represent being filled. So if the representations here are the true analyses of possibly being filled and possibly being empty, one or the other (or some third approach along these lines) will have to be objectively "privileged." But none is; there will always be equally good but incompatible set-theoretic ways to represent various possibilities, just as there are equally good but incompatible set-theoretic candidates for being the number three or for being the ordered pair $<a,b>$.

Some will object that Benacerraf's worry about identifying numbers with sets, and the worry Forrest and Armstrong have about identifying ordered pairs with unordered sets, are really no worries at all. Rather, some will say, it is simply *indeterminate* which of the competing sets is identical with the number three and which unordered set of sets is

[20] "In Defence of Structural Universals," *Australasian Journal of Philosophy*, LXIV (1986): 85–88, here p. 87

$<a,b>$. (There are a variety of ways this sort of reply could go, each invoking a different notion of indeterminacy.[21])

I do not find this reply very plausible. I think it fails to block Benacerraf's objection, and fails to block Armstrong's and Forrest's objection. But my aim here is not to defend Benacerraf or Armstrong or Forrest. And I do not need to. For whatever one thinks of the indeterminacy reply with respect to, say, reducing ordered pairs to unordered sets, one should find it absolutely unacceptable as a way of saving the "objective" version of Heller's abstract counterpart theory.

To begin to see why I say this, consider the following. Faced with, for example, two intuitively similar reductions of ordered pairs to unordered sets, perhaps we could plausibly play the indeterminacy card. But playing that card would not be remotely plausible if faced

[21] Wright develops this reply invoking "Quinean indeterminacy," pp. 125–27. And, as Hazen notes, so would Gilbert Harman and Quine himself—see Hazen's "McGinn's Reply to Wright's Reply to Benacerraf," *Analysis*, XLV (1985): 59–61; Harman's "An Introduction to 'Translation and Meaning', Chapter Two of *Word and Object*," in Donald Davidson and Jaakko Hintikka, eds., *Words and Objections* (Boston: Reidel, 1969), pp 14–26; and Quine's "Replies To Harman" in *Words and Objections*, pp 295–97.

But Wright is wrong (so are Harman and Quine). Marie McGinn raises one objection to Wright, an objection I endorse, in "Wright's Reply to Benacerraf," *Analysis*, XLIV (1984): 69–72. Moreover, suppose that there are more and less natural interpretations of language (cf. Lewis, "New Work for a Theory of Universals," *Australasian Journal of Philosophy*, LXI (1983): 343–77). For example, one might say, it is more natural to interpret 'gavagai' as 'rabbit' than as 'undetached rabbit part' or as any of the other alleged competitors. And so, one might say, 'gavagai' means rabbit. But none of the competing set-theoretic entities are "more natural" referents of the expression 'Merricks's counterparts' than are any of the others. So the sort of thing that would abolish Quinean indeterminacy leaves unresolved the conflict among our incompatible but equally good candidates for being my counterparts. So the problem with objective set-theoretic counterparts is not merely an instance of Quinean indeterminacy.

Someone might respond to my Benacerraf-style objection by saying it merely points out an instance of vagueness due to semantic indecision. But—I reply—in cases of that sort of vagueness, deciding on a single candidate results in added precision. For example, suppose one area with precise borders was designated 'outback'. Then certain sentences would be determinately true that (according to the believer in semantic indecision) were not so before, such as "Ned stepped out of the outback exactly ten seconds ago."

Now suppose the alleged vagueness due to semantic indecision associated with "Merricks's counterparts" were precisified away. This would not add the appropriate sort of precision, in part because what makes the candidates "equally good" is their representing exactly the same possibilities. Because selecting one set-theoretic representation of a possibility from among the various "equally good" candidates cannot deliver the appropriate precision, the problem with "objective" set-theoretic counterparts is not due to vagueness of the semantic indecision sort. Joseph G. Moore defends a similar point while presenting a Benacerraf-style argument against identifying propositions with sets—see "Propositions, Numbers, and the Problem of Arbitrary Identification," *Synthese*, CXX (1999): 229–63.

with the idea that each ordered pair is identified with any and every set that meets or exceeds a certain threshold of structural complexity. And, of course, things get worse if the view in question implied that each ordered pair is also a cat and is a tree and is in fact each and every thing with a sufficiently complex "structure." The moral here would not be that it is indeterminate which of *every complex entity* is identical with $<a,b>$. The moral, obviously enough, would be that it was a mistake to try to reduce that ordered pair to whatever has a sufficiently complex structure.

This is pretty much the place we find ourselves if we say that abstract worlds and their counterparts are sets that represent possibilities in virtue of their "structure," rather than in virtue of how they are interpreted. For although I have gone along with the general framework Heller suggests, there is no reason to think that abstract worlds must really be sets of ordered pairs rather than, say, sets of ordered triples, or quadruples, or...whatever. Indeed, there is no reason—given only the vague notion that "structure" accounts for the relevant sort of representation—to exclude the idea that my being possibly forty feet tall just is the existence of a yawning cat. Yawning cats do, after all, have "structure."

Whatever one makes of the Benacerraf-style objection in other contexts, it seems to me that it is decisive against the Heller-style theory according to which sets represent possibilities only in virtue of their "structure." At the very least, that objection cannot be plausibly blocked by invoking indeterminacy.

The ubiquity of structure is not even the most fundamental problem here. For suppose, just for the sake of argument, we could somehow block the Benacerraf-style objection. The view under discussion here still says that *sets of ordered pairs (of quadruples of numbers and sets of numbers) represent possibilities essentially and intrinsically, in and of themselves, not by way of interpretation.* While there is probably intrinsic representation out there—for example, some mental states are plausibly intrinsically "about" something else and perhaps there are *sui generis* propositions—a set of ordered pairs (of quadruples of numbers and sets of numbers) seems an unlikely place to find it. No one has adequately explained how such sets, in virtue of their "structure," really manage to represent possibilities. Indeed, it seems as if they simply could not do this. No matter what a set's structure, it seems that—if we set aside all interpretation—the set would represent nothing. A set just sits there. After all, in the absence of being appropriately interpreted, the number 1 does not represent *being filled.* And presumably this is true of 1 no matter how the sets are structured that include 1 in their transitive closure.

For bookkeeping purposes, I shall count all of the above worries about objective set-theoretic counterparts as a fourth objection to Heller. For their point is that we cannot save Heller's set-theoretic counterparts by scaling back the role of interpretation.

Objection 5: My fifth and final objection to Heller begins by noting that everyone—including every counterpart theorist—should agree that my possibly being forty feet tall is not analyzed in terms of my being on the same planet as a yawning cat. And talking funny does not help this analysis; it does not help to call the cat "my counterpart" and its yawning "the counterpart of my being forty feet tall." It is obvious that my being possibly forty feet tall is one thing, a cat's yawning another.

The classic objection insists that Humphrey's possibly winning is one thing, an otherworldly twin of his pulling off an upset another. Nevertheless, the defender of "yawning cat" counterpart theory is mistaken if he responds: "I am no worse off than every other counterpart theorist; those who reject my view because it seems (to them) obviously false are just replaying the classic objection." Yawning cat counterpart theory is obviously wrong and we should come right out and say it.

It is obviously wrong. But it is hard to say anything uncontroversial about *why* it is wrong. That is because to do so would be to note some desideratum of an analysis that the yawning cat account fails to satisfy. But, of course, any such desideratum is likely to be controversial. And, more to the point, any substantive "standard for a successful analysis" will be more controversial than the obvious falsity of the claim that my being possibly forty feet tall *just is* my being on the same planet as a yawning cat.

I think Heller is in yawning cat territory. Once we see that counterpart theory is supposed to give us *analyses* of modal properties then we should also see how implausible Heller-style analyses are. Even if we set aside the first four objections, it would still not be plausible that my being possibly forty feet tall *just is* (my existing and) the existence of some set of the sort Heller invokes. Again, my being possibly forty feet tall is not the same thing as there being a particular set of ordered pairs of ordered quadruples of numbers and sets of numbers.

IV. OTHER ABSTRACT COUNTERPART THEORIES

Lewis presents, but of course does not endorse, a version of abstract counterpart theory. (In fact, Heller says his own view is an improved version of the theory Lewis presents.) In sketching a view according to which abstract possible worlds are sets of sentences, Lewis says:

> ...it is not necessary that the worldmaking language should itself be anything like a natural language, or that there should be any way to speak or write its sentences, or that its vocabulary should be finite, or that its sentences should be finite in length. All we need is language in a generalised sense: a system of structures that can be parsed and interpreted. The words can be anything that safe and sane ontology has to offer. They can be individuals that are part of the concrete world, or set-theoretic constructions out of those, or pure sets, or anything else we believe in. It does not matter what they are, if they get their role in representing by arbitrary stipulation. So we could have an infinite number of words: let all the real numbers serve as words, or all the points or regions of spacetime, and stipulate in some wholesale way what they are to mean (*On the Plurality of Words*, *op. cit.*, p. 144).

Here is what Lewis says when—inspired by Jonathan Swift—he considers a "Lagadonian" language in which each thing names itself:

> A language that is interpreted in the Lagadonian way is, nevertheless, interpreted. The ersatzer who presents his worldmaking language must still stipulate how it is interpreted, even if what he stipulates is that it has a Lagadonian interpretation. That is not, of course, the only interpretation the language could bear. It is just an especially easy one to specify. A real number could serve as a name for itself, but could equally well serve as a name for itself plus seventeen. I could be my own name, but alternatively I could be yours; or I could be put to some other use in a language, say as a punctuation mark (*op. cit.*, p. 146).

And here is what Lewis says about our counterparts:

> Just as we take linguistic ersatz worlds as complete descriptions of an entire world, so we should take our linguistic ersatz possible individuals as complete descriptions of parts of the world. We can take them as maximal consistent sets of open sentences of the worldmaking language, with a free pronoun or variable; or, equivalently, as maximal consistent sets of complex predicates. Just as a set of sentences is consistent if they might all be true, so a set of open sentences is consistent if they might all be true *of* something, taken as a value of the common pronoun or variable; and a set of predicates is consistent if something might fall under all of them...(*op. cit.*, p. 149).

Lewis says that the sort of abstract counterpart theory sketched in the quotes above is the "best ersatzism by far" and that he has no "knockdown refutation" of it (*op. cit.*, p. 165). Rather, he notes one or another way in which this theory is more costly or less thorough than his own modal realistic theory. But we should be harsher here than Lewis was. For it should be clear, holding the quotes above up to the light of the earlier arguments against Heller, that the sort of

counterpart theory here on offer is absolutely hopeless (hopeless if taken as providing the true analyses of modal properties). But risking redundancy, I shall make a few comments.

Something is an open sentence, or a complex predicate, because of something we do, because of an interpretation. We play a role in deciding which things are sentences and predicates, and which sentences and predicates those things are. And so, like Heller, the abstract counterpart theorist who analyzes modal properties in terms of sets of open sentences or predicates must say that how we interpret influences the very nature of those analyses. Given one way of interpreting, the analysis of a modal property is partially in terms of one set; given another, the analysis does not involve that set at all.

Like Heller's, the view sketched by Lewis implies that the analysis of modal properties is contingent, since that analysis turns on contingent facts about what—to use Lewis's word—we "stipulate." Also like Heller's, it implies that each and every modal property hangs in limbo—does not even exist—absent the relevant stipulations or interpretations. But, presumably, there are some modal properties that have yet to be accounted for in a properly interpreted world. Given Lewis's version of abstract counterpart theory, those modal properties do not exist. But—as just noted—presumably there *are* such properties. Hence that sort of counterpart theory is wrong.

Multiple "languages" will be equally well suited to the task of being the one out of which the sets are constructed that, in turn, issue in the analyses of modal properties. As a result, multiple but incompatible such analyses of each modal property will be equally good. And so we have a Benacerraf-style objection to abstract counterpart theory of the sort sketched by Lewis, if we should try to avoid the first three objections by downplaying the role of interpretation in that theory.

More importantly, if we deny interpretation a role here, it is hard to see how we have anything at all like a language out of which to construct the relevant sets. It is hard to see how we could have *any* good analyses, much less—as the Benacerraf objection requires—too many. Avoiding the first three objections by asserting that sets represent objectively is swallowing a frog to get the fly. (The fifth objection—the claim that my being possibly forty feet tall is not the same thing as there being a set or other abstract object—cannot be avoided at all.)

Perhaps someone will reply that the Lagadonian interpretation (or something like it) *is* privileged. This might be the reply, for example, of someone who thinks a singular proposition has its subject as a "constituent."[22] There are a variety of ways this might go. Such a view

[22] Cf. Bertrand Russell, "On Denoting," *Mind*, XIV (1905): 479–93.

might claim, for example, that the proposition that Merricks is forty feet tall just is the ordered pair <Merricks, *being forty feet tall*>. Now I do not think this ordered pair represents anything at all. It just sits there. And if it did, all by itself, manage to represent something, I do not see why it would represent my being forty feet tall rather than, for example, my not being forty feet tall. Indeed, I do not even see why, when a member of such a pair, I am more fit to represent myself than, say, the number seventeen. So I do not think the Lagadonian approach helps.

But I am willing to concede, for the sake of argument, that *if* one were to defend objective representation by sets, the least implausible version would be broadly Lagadonian. Even if this approach could be made to work, however, it would be of no help to abstract counterpart theory. For, according to this approach, an object represents itself in worlds according to which it exists; it is somehow a constituent of every world that represents it; it is, so to speak, its own counterpart in various worlds. Given this view, the "counterpart relation" is numerical identity. But the counterpart relation is not identity. So this view is not a version of counterpart theory at all. Indeed, because this view says that each object represents itself in various worlds, its closest cousin is the paradigm of anti-counterpart theory: modal realism with overlap.[23]

(Alternately, look at it this way. A reduction of modal properties to abstract worlds is counterpart-theoretic if and only if (or perhaps to the extent that) it delivers the relevant theoretical payoff: sortal-relative counterpart relations, intransitive "transworld identity," and some sense to the idea of contingent identity.[24] So-called Lagadonian counterpart theory delivers none of this because its "counterpart relation"—numerical identity—is not sortal-relative, is transitive, and is in no way contingent.)

My objections to Heller's abstract counterpart theory are objections to abstract counterpart theory as sketched by Lewis. Moreover, I think those objections are—or can easily be transformed into—objections to any abstract counterpart theory. For our "counterparts," if they are abstract, represent us only in virtue of being thus interpreted or in

[23] See Lewis, *On the Plurality of Worlds*, pp 198–209.

[24] Counterpart theory's bearing on contingent identity and the sortal relativity of *de re* modal properties is widely discussed (see, for example, Lewis, *On the Plurality of Worlds*, pp. 248–63). Somewhat less familiar is its handling of "Chisholm's paradox" (see Chisholm, "Identity through Possible Worlds: Some Questions," *Noûs*, 1 (1967): 1–8). Here I note only that counterpart-theoretic solutions to that paradox trade on the intransitivity of the counterpart relation (see, for example, Lewis, *On the Plurality of Worlds*, pp. 243–46, and Forbes).

virtue of what we stipulate. (Thus, the first three objections to Heller can be turned against any abstract counterpart theory.) Relatedly, an interpretation or stipulation is required because abstract structures do not seem suited to represent all on their own. And if some were thus suited—say in virtue of their "structure"—others would be as well, and would be so in such a way as to yield mutually inconsistent analyses of modal properties (thus the fourth objection). Finally, any kind of abstract counterpart theory will force us to identify the having of a modal property by an object with the existence of (that object and) an abstractum (thus the fifth).[25]

Rather than reject abstract counterpart theory in light of these objections, the devoted counterpart theorist may take them as a recipe for a defensible counterpart theory. That is, she may say that abstract counterpart theory is true, but then add that it lacks all the features the above objections trade on. So suppose she says the following. Abstract counterpart theory is true. But there is a single correct abstract analysis for each modal property, an analysis that involves abstracta having their representational features essentially and independent of any interpreting or stipulating on our part. (She thereby hopes to block my first three objections to abstract counterpart theory.) And to sidestep all the worries contained in my fourth objection, she denies that worlds are sets and she denies that worlds represent possibilities in virtue of their "structure." Rather, she insists, a world represents various possibilities just because it does; and other abstracta fail to represent those same possibilities just because they do. Worlds, she says, are *sui generis* propositions (or states of affairs, and so forth).

This reply is the last hope for abstract counterpart theory, the only way to avoid the above four objections. (The abstract counterpart theorist cannot possibly avoid the fifth, although she can question its force.) Those who find *sui generis* propositions unacceptable—either because of Lewis's criticisms of "magical ersatzism" or for other reasons[26]—will likewise refuse to accept the above reply. And so some will think this response fails to save counterpart theory simply because what it says about representation is not credible.

But I do not want to leave it at that. This is because, as Peter van

[25] Some might claim that the abstract counterpart theorist is no worse off with respect to the fifth objection than is the Platonist who thinks that *being possibly forty feet tall* is an abstract object. But that claim is mistaken. For the Platonist does not say that my existing along with that abstract object is what it is for me to be possibly forty feet tall. Rather, she says that my *exemplifying* that object is what it is for me to be possibly forty feet tall. We might wonder what exemplification is and whether the Platonist's story is plausible; but that is a different topic.

[26] For Lewis's criticisms, see *On the Plurality of Worlds*, pp. 174–91.

Inwagen[27] has argued, we may be able to resist Lewis's attack on magical ersatzism. And, Lewis's attack notwithstanding, *sui generis* propositions are a familiar tool in philosophy, one that I do not want to reject out of hand. Nevertheless, I shall argue, such propositions (or states of affairs or other "magical" representers) do not provide a response to my objections that the counterpart theorist can endorse. For even though some other believers in abstract possible worlds may embrace magical ersatzism, I shall argue that the counterpart theorist in particular cannot. But I cannot develop that argument until I explore some of what separates abstract counterpart theorists from other believers in abstract worlds. This is the project of the next two sections.

V. ABSTRACT WORLDS WITHOUT ANALYSIS

My arguments against abstract counterpart theory do not show that Lewis's theory is false. For, given his modal realism, counterpart relations hold and are what constitute modal properties independent of any interpretations on our part. After all, on Lewis's view my counterparts are similar to me in the most objective and straightforward way possible; they likewise straightforwardly have the properties I possibly have.[28]

Indeed, one might suspect that only Lewis's theory is safe from the above arguments. One might suspect that those arguments trouble any non-Lewisian approach to *de re* modality—or at least any approach invoking abstract possible worlds. For example, just so long as modal properties are analyzed in terms of interpreted abstract structures (whether the analysis is counterpart-theoretic or otherwise) we face analogues of my first three objections to Heller's theory.

Moreover—setting aside *sui generis* propositions for the moment—the fourth objection raises problems for abstract worlds generally,

[27] See "Two Concepts of Possible Worlds," *Midwest Studies in Philosophy*, XI (1986): 185–213. Objecting to Lewis's pejorative label, van Inwagen renames the position "Unsound Abstractionism," 'Unsound' being an acronym for Unscientific Naive Superstitious Obscurantist Unenlightened Neanderthal Dogmatic.

[28] Lewis does say that modal *predicates* are context relative. But, for Lewis, it is wholly objective whether an object stands in the various relations that constitute sortal-relative modal properties.

Because of the arguments in this paper against linguistic ersatzism, I think Lewis's theory is most defensible if representation drops out. That is, his view is most defensible if we deny that it is my counterpart's representing me as being forty feet tall that does the work; rather, it's being forty feet tall and otherwise appropriately similar to me does it. We have here the "reduction without representation" that I criticized in section III as being unmotivated; but in Lewis's case, I think it is motivated. Given his ontology, his choice of counterparts does not seem arbitrary.

counterpart-theoretic or otherwise, that try to avoid the first three objections by abandoning interpretation. For worlds seem unable to represent anything at all without interpretation lending a hand. Without our interpreting, it is hard to see how anything could be true *according to* a world. And if worlds did manage to represent on their own—we can play along by saying they do so in terms of their "structure"—they would run afoul of the Benacerraf-style objection.

Consider also this objection. Believers in abstract possible worlds think that I am possibly forty feet tall just in case, in some abstract world, I am forty feet tall. But my being possibly forty feet tall is *not* the same thing as there being some abstract world—some set or some other abstractum—according to which I have that property. Recalling my final objection to abstract counterpart theory: it is not plausible that my being possibly forty feet tall *just is* the existence of an abstract whatnot.

The objections just raised presuppose that the noncounterpart theorist tries to *analyze* modal properties (or possibility and necessity) in terms of abstract possible worlds. And if that presupposition were correct, I think the objections would be decisive. But it is not correct; for the noncounterpart theorist does not try to analyze modal properties, at least not typically.[29] To illustrate this, consider a typical approach (along the lines of Plantinga's *The Nature of Necessity*). Start with my being possibly forty feet tall. Since I have this property, at least some abstracta according to which I am forty feet tall are possible. At least some of those abstracta are "maximal"; that is, at least some of them either include or preclude every other way things could be. Call the maximal ones "possible worlds." Conclusion: there is a possible world in which I am forty feet tall.

The above reasoning is controversial. Some will deny the existence of representing abstracta. Others will believe in such abstracta, but deny that any are "maximal." And yet others will accept the existence of maximal representing abstracta, but will object to calling them "possible worlds." Happily, we can ignore these and similar controversies. For the point here is not that the view just outlined is true and felicitous. Rather, the point here is that that view—true or not—has nothing to do with analyzing modal properties.

Nor do other typical noncounterpart-theoretic accounts of abstract possible worlds. As noted, the above objections to noncounterpart-theoretic abstract possible worlds assume, falsely, that such worlds are in the business of analyzing modality. Those objections succeed only

[29] Cf. Plantinga, p. 44, Kripke, p. 19, note 18; and van Inwagen, p. 187.

given the false assumption. And so standard noncounterpart-theoretic accounts of abstract possible worlds are immune to my objections to abstract counterpart theory.

Let me say just a bit more to justify this last claim. As already noted, the standard sort of noncounterpart theory *starts* with modal properties. It then analyzes worlds in terms of them. Therefore only the confused will object that there being a possible world in which I am forty feet tall is irrelevant to my being possibly forty feet tall. Moreover, no matter how one might develop this position with respect to the role of interpretation, the Benacerraf-style worry about there being too many "abstract world" analyses of what it is for me to be possibly forty feet tall never gets going. For there are not too many analyses—there is not even one! Analysis of modal properties is not the topic here. So there is no worry that the nature of modal properties—or the modal status of their nature or the existence of the properties themselves—is held hostage to, or somehow compromised by, our interpreting. Even if our interpretations play a role in what abstract worlds represent, such interpretation would not bear on the analysis or nature of modal properties.

Abstract counterpart theory, if a true theory about modality, must offer counterpart-theoretic analyses of modal properties. But, I have argued, all abstract world analyses fail, including of course the counterpart-theoretic ones. So—if those arguments are correct—abstract counterpart theory is not true. My arguments against that theory do not threaten Lewis's nonabstract counterpart theory; nor do they trouble the typical noncounterpart theorist who believes in abstract possible worlds.

VI. SUI GENERIS PROPOSITIONS AND THE ANALYSIS OF MODAL PROPERTIES

As noted, the typical noncounterpart theorist does not try to analyze modal properties in terms of abstract worlds. Thus she is immune to my arguments against abstract counterpart theory. But, I said, if she tried to analyze modal properties in terms of abstract possible worlds, she would not be immune. And so I argued that all analyses of modal properties in terms of abstract worlds fail.

In arguing against all such analyses or reductions, I set aside the idea that abstract worlds were *sui generis* propositions (or states of affairs or the worlds of magical ersatzism). I now want to ask whether that idea offers a way of salvaging possible worlds analyses of modal properties in general and the counterpart-theoretic species of such analyses in particular.

Let us begin by imagining two philosophers who agree that possible

worlds are maximal and possibly true *sui generis* propositions. And they agree that a person "exists in a possible world" if and only if the world represents the person as existing (and having various other properties). But they disagree about something. One reduces modal properties to representation by worlds, the other not.

Given the extent to which our two philosophers agree, one might suspect that their disagreement comes to nothing. But consider how each would understand the claim that I am forty feet tall in possible world β. The nonreductionist would understand this as meaning: if β were true, then I would be forty feet tall. In contrast, the reductionist (and so the analytic counterpart theorist) understands this to mean: β represents me as being forty feet tall.[30]

With this in mind, consider *sui generis* propositions. Given these propositions—given magical ersatzism—there is no explanation of how representation *works*. But the magician who does not reduce modal properties to representations can help us understand something of what representation *is*. She can say that what it is for me to be represented in a world as being forty feet tall is the following: if things were as that world represents them, then I would be forty feet

[30] The nonreductionist's view cannot be a version of counterpart theory. But Robert Stalnaker has considered a (so-called-by-him) version of abstract "counterpart theory" that, in the relevant respects, is like the nonreductionist's view. After presenting his theory, Stalnaker entertains this objection "Still, the critic may respond, what's the point? This theory will surely be just a clumsy notational variation of the standard theory [that is, the noncounterpart-theoretic account of abstract possible worlds]" (p. 131). I salute the critic and endorse the criticism. Stalnaker replies that his theory can make sense of contingent identity in the way that counterpart theory can and in a way that noncounterpart theory cannot The critic is answered if and only if Stalnaker's reply is correct.

I am one object. Now suppose two different things represent me in some world This might seem to imply that I am possibly two things. So it might seem to imply that two things are both actually me only contingently. (This is how Stalnaker thinks his view accommodates contingent identity.)

But according to Stalnaker, a bit of abstracta's representing me as being a certain way means that, were the world in which it is included actual, then I myself would be that way. (This characterization of representation in a world is, I say, enough to render a view noncounterpart-theoretic.) Therefore, even though two distinct bits of abstracta represent me I am not—given Stalnaker's constraint on representation just noted—represented as being two things For they both represent me as being a certain way only if, were that world actual, I—one thing—would be that way But if that world were actual, I would be one thing, not two. So in that world—given Stalnaker's constraint on representation and the assumption that the world is indeed possible—I am represented as one thing even if I have more than one representation

Thus Stalnaker cannot answer the critic. If Stalnaker amended his view by claiming that for me to be possibly two people *just is* (is reduced to or analyzed as) my having two representations (*qua* person) in a single world, he would be able to account for contingent identity. But if he makes this move, he becomes a garden-variety abstract counterpart theorist, analyzing possibilities in terms of representations

tall. More generally, what it is for any object *O* to be represented in a world as having any property *F* is the following: if things were as that world represents them, *O* would be *F*.

The reductionist cannot explain representation by worlds in this way. For, as we have already seen, she reduces *if things were as a world represents them, O would be F* to *a world represents O as being F*. Given reductionism, the latter cannot be explained in terms of the former; it is rather the other way around. So when we ask someone who reduces modal properties to abstracta to explain what it is for a world to represent *O* as being *F*, she must—if she believes a world is a *sui generis* proposition—say nothing. On her view, representation by a world is completely mysterious.

Freed of reductive ambitions, the magical ersatzist can explain representation by a possible world in terms of how things would be, were they as they are represented. Freed of magical ersatzism, one who reduces modal properties to representation by abstract worlds can explain that representation in terms of interpretation and language. But the combination of magical ersatzism and the reduction of modal properties to what is represented by worlds is fatal. For it renders representation by a world—and hence modal properties reduced to such representation—unintelligible.

The argument just given takes aim at any attempt to reduce *de re* modality to abstract representations. So the point of the above argument is that anyone who wishes to reduce *de re* modality to representation by abstract worlds cannot rely on magical ersatzism. And of course this implies that the abstract counterpart theorist cannot rely on magical ersatzism. As we have seen from the start of this article, abstract counterpart theory has analytic ambitions. It aims to analyze (or reduce) modal properties to representations in worlds by counterparts.[31]

[31] This point should already be familiar given earlier discussions and arguments. Let me note, finally, that we find an explicit statement of it in Lewis:

> According to [counterpart-theoretic] ersatz modal realism, there are abstract surrogates fit to play the same theoretical roles as the concrete *possibilia* that are to be rejected. These are abstract entities capable, somehow, of representing concrete entities. They represent in a double sense. (1) They are representations, so that it somehow makes sense to speak of what is the case *according to* them; and thereby (2) they are representatives, taking the place of what they purport to represent (*On the Plurality of Worlds*, p. 137).

The nonreductionist who believes in abstract possible worlds will join the reductionist in endorsing Lewis's claim (1); both say that we are represented in abstract worlds. But they part ways at claim (2), the nonreductionist denies that the existence of such a representation just is—or, in Lewis's words, "takes the place of"—my having the relevant modal property.

There is a second problem with the combination of counterpart theory and magical ersatzism, a problem that afflicts the counterpart-theoretic species of reductionism in particular. The magical ersatzist believes in *sui generis* propositions (or states of affairs, and so forth). So suppose that we have *sui generis* propositions. And add that an object exemplifying properties P_1 through P_n is "appropriately like me." The sort of counterpart theory we are here considering would then claim, roughly, that my being possibly forty feet tall is reduced to the *de dicto* possibility of the proposition that there is some object exemplifying properties P_1 through P_n that is forty feet tall. (And this proposition must somehow be part of a "possible world." For something "appropriately like me" is my counterpart only if it is the thing relevantly most like me in a given world.)

We are supposing that there are *sui generis* propositions. But if so, surely some are singular. If there is the proposition that there is some object exemplifying P_1 through P_n that is forty feet tall, presumably there is also the proposition that Merricks is forty feet tall. Now ask yourself which of the following is the more natural or obvious or intuitive or plausible *de dicto* reduction of my being possibly forty feet tall:

(i) Possibly, Merricks is forty feet tall.
(ii) Possibly, some object exemplifying P_1 through P_n [that is, some object appropriately like Merricks] is forty feet tall.

Clearly, (i) is.

This does not beg any questions against abstract counterpart theory. The goal, we are imagining, is to reduce or analyze each *de re* modal property in terms of a *de dicto* modality. We accomplish this by analyzing my being possibly forty feet tall in terms of either (i) or (ii). The question then—at this stage in the dialectic—is, given (i) and (ii), which better captures the idea that I am possibly forty feet tall. The answer is obviously (i); no *de dicto* modality could ever do better.

But the counterpart theorist cannot reduce the claim that I am possibly forty feet tall to the *de dicto* possibility of the proposition that Merricks is forty feet tall. She cannot do this because, obviously enough, the resulting account of *de re* modality is not counterpart-

---

As indicated by the bracketed insertion, I take Lewis's comments to be about abstract *counterpart theory*. I so take them because if abstracta play the role of his concrete possible individuals, then abstracta play the role of—and so are—counterparts. Yet Lewis might object, on grounds of terminology, to my bracketed insertion. For the way Lewis uses 'counterpart' in much of *On the Plurality of Worlds* seems to permit, just by definition, only modal realistic counterparts (see, for example, pp. 195–96). Late in the book, however, he treats ersatz counterpart theory as a coherent proposal (pp. 237–39).

theoretic. For example, this account, like the view that each object represents itself in the worlds in which it exists (see section IV), fails to deliver the theoretical flexibility of counterpart theory.

(An additional small point in favor of the noncounterpart-theoretic approach to reducing the *de re* to the *de dicto* is that it is the more economical. As already noted, counterpart theory requires worlds. But the noncounterpart theorist needs only singular propositions, like that in (i) above.)

Abstract counterpart theory must embrace magical ersatzism to avoid the first four objections developed earlier in the paper, objections presented in their original form against Heller's theory. (The fifth objection is unavoidable.) But magical ersatzism's *sui generis* propositions undermine abstract counterpart theory. They do so, as argued above, for two reasons. First, any abstract *de re* modal reductionism (which includes all versions of abstract counterpart theory) renders the relevant notion of representation unintelligible. Second, if we allow *sui generis* propositions, we should allow singular *sui generis* propositions. But the latter provide a noncounterpart-theoretic *de dicto* reduction of each *de re* modal property, a reduction which trumps that suggested by the counterpart theorist.

As we have just seen, the counterpart theorist cannot embrace magical ersatzism. But, as we saw earlier (section IV), only magical ersatzism offers a hope of blocking (four of) the objections to abstract counterpart theory, objections to any possible abstract counterpart-theoretic analysis of a modal property. So those four objections cannot be blocked; neither can the fifth. I conclude that abstract counterpart theory is false.

VII. MERELY HEURISTIC COUNTERPART THEORY

I simply assume that Lewis's ontology of multiple universes is not true. So I conclude that his counterpart theory fails. And, I have argued, abstract, non-Lewisian, counterpart theories also fail. So counterpart theory of every variety fails. That is, counterpart theory does not give us the true account of the nature of *de re* modal properties.[32]

[32] I have argued against *modal* counterpart theory Sider defends *temporal* counterpart theory, offering counterpart-theoretic analyses of an object's past- and future-rooted properties—see his *Four-Dimensionalism*, pp. 188ff, and his "All the World's a Stage," *Australasian Journal of Philosophy*, LXXIV (1996) 433–53. (Katherine Hawley defends a similar view in *How Things Persist* (New York: Oxford, 2001).)

Sider presupposes that the past and the future are "just as real" as the present. This allows him to suppose that an object's temporal counterparts are kindmates of the object itself. (In this way, Sider's temporal counterpart theory parallels modal-realistic counterpart theory.) As a result, my arguments here against abstract (modal) counterpart theory do not threaten Sider's version of temporal counterpart theory. If, however, a presentist—someone who denies "the reality of past and future"—were

But perhaps some self-styled counterpart theorists offer no such account. Perhaps they employ "counterpart theory" not to reveal the metaphysics of modality, but rather as a heuristic to help us imagine, or to help them present, possibilities in a supposedly illuminating way. I suppose, for example, that most philosophers who use possible worlds as a mere heuristic have in mind worlds of Lewis's sort; they will almost inevitably end up with merely heuristic counterpart theory.[33]

Counterpart theory, if the true theory of modality, delivers novel modal results. These results—and the ways in which they solve certain puzzles—make counterpart theory attractive to some.[34] But the very features that make counterpart theory attractive as a theory of the nature of modality make it a dismal heuristic.

Here is one example. I can easily picture a "world in which I have two (personal) counterparts." This picture suggests that there is some sense to be made of the claim that there is a possible world in which I am two people, and so some sense to be made of the claim that, possibly, I am two people. But this suggestion is totally misleading. For—if analytic counterpart theory is false—to say that there is a world in which I am two people is to say that there is a way things could be such that, if things were that way, I myself would be two people. But if two people were identical with me, each would be identical with one person: me. Absurd.[35]

---

to endorse temporal counterpart theory, then temporal counterparts would have to be "abstract", and presentist temporal counterpart theory would be undermined by the arguments of this article

[33] Gideon Rosen considers something like this when he introduces a nonanalytic version of modal fictionalism—"Modal Fictionalism," *Mind*, XCIX (1990): 327–54, see p. 354. "Heuristic fictionalism" should be rejected as misleading since, for example, it supports false claims like 'I might have been two people' on the grounds that, according to Lewis's theory, I might have been two people. The stronger, analytic version of fictionalism—the version with which Rosen is primarily concerned—is subject to basically the same objections as abstract counterpart theory. Moreover, there are worries that Rosen's suggested analyses arbitrarily choose Lewis's story (as opposed to, for example, Armstrong's) (see p. 332, note 10).

[34] See, for example, Stalnaker, pp. 124–26; and Sider, *Four-Dimensionalism*, pp 206–07.

[35] Lewis makes a similar point, *On the Plurality of Worlds*, pp 253–54. Relatedly, Sider presents an argument for the necessity of distinctness that, he says, is sound unless counterpart theory is true or modal predicates equivocal (*Four-Dimensionalism*, pp. 223–24) I think that there are *de re* modal properties. So I say that if modal predicates are appropriately equivocal, modal properties are sortal-relative, but this makes sense, I would argue, only if such properties are analyzed in terms of sortal-relative counterpart relations (cf. Harold Noonan, "Indeterminate Identity, Contingent Identity and Abelardian Predicates," *Philosophical Quarterly*, XLI (1991): 183–93, p. 190, note 3). (Lewis says that (except for modal realism with overlap) rival theories do as well as his at making sense of the "inconstancy of modal predicates." But note that these rivals are all versions of (abstract) counterpart theory (see Lewis, *On the Plurality of Worlds*, pp. 259–61).)

A similar point could be made about heuristic counterpart theory's suggesting that "transworld identity" is intransitive. Or about its suggesting that "transworld identity" and (as a result) modal properties are sortal-relative. The moral in all these cases is the same. Heuristic counterpart theory suggests that we think of modality as if counterpart theory were true. But it is not true. And so we have no right to think of modality in that way.

Counterpart theory is not just a less prosaic, more picturesque way of saying what noncounterpart theorists had been saying all along.[36] And so thinking in terms of counterpart theory is useful if and only if it is an accurate way to think of modal properties; that is, if and only if there are counterpart-theoretic analyses of modal properties. But because analytic counterpart theory is false—because it does not reveal the true nature of *de re* modality—heuristic counterpart theory is positively misleading.

TRENTON MERRICKS

University of Virginia

[36] Cf. Lewis, "Counterpart Theory and Quantified Modal Logic," this JOURNAL, LXV, 5 (March 7, 1968) 113–26, here p. 115.

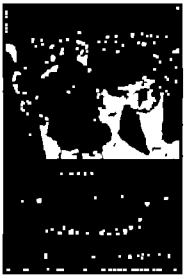
NOTES AND NEWS

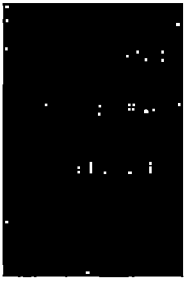
The Editors are saddened to report the death of the brilliant philosopher Sir Bernard Arthur Owen Williams, on June 10, 2003, at the age of 73. Born in Westcliff, Essex, England in 1929, Williams attended Balliol College, Oxford University, and was appointed a Fellow at All Souls College, Oxford. He took a post at New College from 1954–59, moved on to University College, London during the years 1959–64, and then on to (now defunct) Bedford College. In 1967, Williams was appointed Knightsbridge Professor of Moral Philosophy at Cambridge University, and from 1979–87, he served as the provost of Kings College. From there, he went to the University of California at Berkeley, becoming Monroe Deutsch professor of philosophy. Williams also held the White's Chair of Moral Philosophy at Oxford from 1990–96, and after his retirement became once again a Fellow at All Souls College.

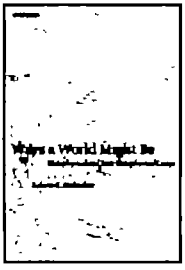
Williams received many awards, fellowships, honorary degrees, and visiting professorships throughout his career. He was knighted in 1999 and sat on several public commissions. Williams's publications include *Morality: An Introduction to Ethics* (1972); *Problems of the Self* (1973); *Utilitarianism: For and Against*, a book co-authored with J.J.C. Smart (1973); *Descartes: The Project of Pure Enquiry* (1978); the anthology *Utilitarianism and Beyond*, co-edited with Amartya Sen (1982); *Moral Luck* (1981); *Ethics and the Limits of Philosophy* (1985); *Shame and Necessity* (1993); *Making Sense of Humanity* (1995); *Plato* (1998); and his final work *Truth and Truthfulness: An Essay in Genealogy* (2002).

Such a catalogue of achievements does little to convey the astounding philosophical personality of the man, a personality larger and more vital than any professional understanding of the term 'philosopher'. He was often and perhaps rightly described as the "cleverest" person in the subject. The almost ceaseless flow of wit, the unaffected and unstuffy public persona, and the effortless capacity as well as the urge to lace even the most serious and penetrating of his remarks with humor, even frivolity, brought to philosophy a quite unusual aspect of fun while he was alive and doing it. It will be the more dull and the more routine, now that he is not doing it. This irreverence was not merely a matter of personality, however. Quite apart from the genuine illumination and unfailing stimulus of his many books and articles on a very wide range of themes, his greatest overall contribution to the subject lay in the highly effective and lasting skepticism he visited upon the entire theoretical tradition in ethics—its doctrinal claims as well as its aspirations to achieve an unattainable strictness of method and principle. In politics, too, Williams could not present his deeply liberal and progressive views without the occasional laugh or smack at their own tendency to piety and superiority. The influence of Nietzsche was visible in these qualities and he avowed that influence explicitly in his work in the last decade or more of his life; but they were not merely a product of scholarly influences. They were native to the mind and the man.

New from
Oxford







Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556
Visit our website at www.oup-usa.org

OXFORD

Interpreting Kant's Critiques
KARL AMERIKS
Karl Ameriks here collects his most important essays to provide a uniquely detailed and up-to-date analysis of Kant's main arguments in all three major areas of his work: theoretical philosophy (*Critique of Pure Reason*), practical philosophy (*Critique of Practical Reason*), and aesthetics (*Critique of Judgment*).
October 2003 paper $22.00 cloth $65.00

The Meanings of Social Life
A Cultural Sociology
JEFFREY C. ALEXANDER
"Boldly, often brashly challenging almost every cultural analyst in sight, Jeffrey Alexander here states and illustrates his strong program for analysis of culture as a coherent, autonomous social realm. The Holocaust, Watergate, computers, and contemporary American society at large all provide starting points for Alexander's distinctive reflections on social experience."
—Viviana A. Zelizer, author of *The Social Meaning of Money*
2003 $29.95

Sense, Reference, and Philosophy
JERROLD J. KATZ
Sense, Reference, and Philosophy develops the far-reaching consequences for philosophy of adopting non-Fregean intensionalism, showing that long-standing problems in the philosophy of language, and indeed other areas that appeared intractable, can now be solved. Katz proceeds to examine some of those problems in this new light, including the problem of names, natural kind terms, the Liar Paradox, the distinction between logical and extra-logical vocabulary, and the Raven paradox.
October 2003 $49.95

Ways a World Might Be
Metaphysical and Anti-Metaphysical Essays
ROBERT C. STALNAKER
This volume begins with reflections on the general idea of a possible world, and then uses the framework of possible worlds to formulate and clarify some questions about properties and individuals, reference, thought, and experience. The essays also reflect on the nature of metaphysics, and on the relation between questions about what there is and questions about how we talk and think about what there is.
October 2003 paper $22.00 cloth $72.00

Agency and Self-Awareness
Issues in Philosophy and Psychology
EDITED BY JOHANNES ROESSLER AND NAOMI EILAN
In recent years there has been much psychological and neurological work purporting to show that consciousness and self-awareness play no role in causing actions, and indeed to demonstrate that free will is an illusion. The essays in this volume subject the assumptions that motivate such claims to sustained interdisciplinary scrutiny.
October 2003 paper $27.50 cloth $77.00

Important Questions. Thoughtful Answers.



Philosophy is often criticized for being irrelevant to social issues of wide public concern. Not *Philosophy & Public Affairs*! In every issue it demonstrates the relevance of contemporary philosophical thought. It is a journal in which philosophers with different viewpoints, along with philosophically inclined writers from many academic disciplines, bring their distinctive methods to bear on problems that concern all readers—scholars and citizens alike.

THE BEST IN CONTEMPORARY THOUGHT

Philosophy & Public Affairs presents important work by today's most thoughtful writers

Recent articles include:

Samuel Freeman, "Illiberal Libertarians: Why Libertarianism Is Not a Liberal View"
Barbara Herman, "The Scope of Moral Requirement"
Samuel Scheffler, "What Is Egalitarianism?"
Edna Ullmann-Margalit and **Cass R. Sunstein**, "Inequality and Indignation"
Bernard Williams, "From Freedom to Liberty: The Construction of a Political Value"

Don't miss out. Subscribe today.

Philosophy & Public Affairs is published quarterly by Princeton University Press and distributed by Blackwell Publishing.

2003 Annual Subscription Rates (U.S.) Individuals, $40.00; Institutions, $80.00; Students, $20.00. Canada and Mexico, add $5.00. Outside North America, add $8.50

How to Order: Journal Customer Services
Blackwell Publishing
350 Main Street, Malden, MA 02148
Phone: 800-835-6770 • Fax 781-388-8232

Philosophy & Public Affairs: The Best in Contemporary Thought

HB Humanity Books

SARTRE AND CAMUS

A Historic Confrontation

Edited and Translated by David A. Sprintzen and Adrian van den Hoven



In 1952, Jean-Paul Sartre engaged Albert Camus in a celebrated and bitter public confrontation that had wide-ranging cultural significance. The year before, Camus had challenged the prevailing political wisdom in his renowned work, *The Rebel*. In response he was attacked in print, first by Francis Jeanson writing in *Les Temps Modernes*, a journal edited by Sartre, and then by Sartre himself. In a series of highly publicized articles, these literary and cultural titans locked horns over human values, social and political policy, the nature of human freedom, the meaning of history, and the future of Western civilization.

Readers will not only be drawn into the issues raised by these two great thinkers but realize that their debate is still with us, perhaps more forcefully than ever.

275 pp • ISBN 1-59102-157-X • HC $45

GLOBALIZATION AND JUSTICE

Kai Nielsen

Philosopher Kai Nielsen persuasively argues that capitalist globalization is the means by which already wealthy interests seek to aggrandize their power and control at the expense of citizens in poor nations. Nielsen contends that democratic socialism is still humanity's best hope for achieving a classless, nonracist, and nonsexist world community.

377 pp • ISBN 1-59102-054-9 • HC $39

NATURALIZING THE TRANSCENDENTAL

A Pragmatic View

Sami Pihlström

In this powerful and original work, Finnish philosopher Sami Pihlström argues that the transcendental and the pragmatist traditions should converge, or at least supplement each other, instead of being regarded as rivals. He proposes a workable middle ground between extreme realism on the one hand and extreme postmodernist skepticism and relativism on the other.

390 pp • ISBN 1-59102-104-9 • HC $49

A HOUSE DIVIDED

Comparing Analytic and Continental Philosophy

Edited by C. G. Prado

The aim of this collection is to reconsider the often facile characterization of major thinkers as belonging to either one or the other philosophical tradition.

329 pp • ISBN 1-59102-105-7 • PB $35

For exam copies contact LRisio@prometheusbooks.com

Call toll free (800) 421-0351 • Fax (716) 691-0137 • www.prometheusbooks.com

University Library
AIIPUR CAMPUS

OUTSTANDING SCHOLARSHIP FROM CAMBRIDGE

The Cambridge History of Philosophy 1870-1945
Edited by Thomas Baldwin
0-521-59104-X Hardback $120 00*

The Innovations of Idealism
Rudiger Bubner
Modern European Philosophy
0-521-66262-1 Hardback $55 00

The Science of the Struggle for Existence
On the Foundations of Ecology
Gregory J. Cooper
Cambridge Studies in Philosophy and Biology
0-521-80432-9 Hardback $60 00

The Cambridge Companion to Medieval Jewish Philosophy
Edited by Daniel H. Frank and Oliver Leaman
Cambridge Companions to Philosophy
0-521-65207-3 Hardback $75 00
0-521-65574-9 Paperback $27.00

Norms in a Wired World
Steven A. Hetcher
Cambridge Studies in Philosophy and Law
0-521-45436-0 Hardback $65 00*

Pursuing Equal Opportunities
The Theory and Practice of Egalitarian Justice
Lesley A. Jacobs
Cambridge Studies in Philosophy and Public Policy
0-521-82320-X Hardback $65 00
0-521-53021-0 Paperback $24 00

The Value of Knowledge and the Pursuit of Understanding
Jonathan L. Kvanvig
Cambridge Studies in Philosophy
0-521-82713-2 Hardback $60 00

The Cambridge Companion to Medieval Philosophy
Edited by A.S. McGrade
Cambridge Companions to Philosophy
0-521-80603-8 Hardback $60 00
0-521-00063-7 Paperback $24 00

An Introduction to Formal Logic
Peter Smith
0-521-81033-7 Hardback $55 00*
0-521-00804-2 Paperback $20 00*

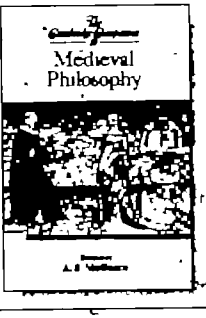
Kierkegaard's Relations to Hegel Reconsidered
Jon Stewart
Modern European Philosophy
0-521-82838-4 Hardback $55 00

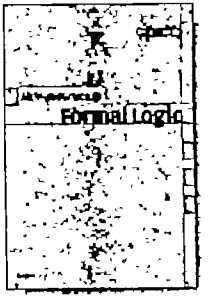
Consent to Sexual Relations
Alan Wertheimer
0-521-82926-7 Hardback $70 00
0-521-53611-1 Paperback $26 00

Democratic Distributive Justice
Ross Zucker
0-521-53355-4 Paperback $23 00

*Prices subject to change







CAMBRIDGE UNIVERSITY PRESS 800-872-7423 us.cambridge.org/philosophy

THE JOURNAL OF PHILOSOPHY

Calcutta University Library
ALIPORE CAMPUS

VOLUME C, NUMBER 11
NOVEMBER 2003

page

553 *Pyrrhic Victories for Scientific Realism* P. Kyle Stanford

573 *Goodman's Problem and Scientific Methodology* Peter Godfrey-Smith

591 *The Metaphysics of Realization, Multiple Realizability, and the Special Sciences* Carl Gillett

604 NEW BOOKS: PAPERBACK REPRINTS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

You may also visit our website at www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2003

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset and printed by Capital City Press, Montpelier, VT.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U S Copyright Law. This consent does not extend to any other kinds of copying More information on our terms may be obtained by consulting our January issue or by writing to us.

POSTMASTER Periodical postage paid at New York, NY, and other mailing offices

POSTMASTER Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027

© Copyright 2003 by the Journal of Philosophy, Inc.

ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY
VOLUME C, NO. 11, NOVEMBER 2003

Calcutta University Library
ALIPUR CAMPUS

PYRRHIC VICTORIES FOR SCIENTIFIC REALISM*

Scientific realists suggest we have good reasons to believe that our best current scientific theories offer us literally true (or probably true, or approximately true, or probably, approximately true) descriptions of how things stand in otherwise inaccessible domains of nature. Their most influential argument for this position is the abductive, explanationist, or "miracle" argument: in Hilary Putnam's memorable phrase, that the success of science would be a miracle if our scientific theories were not true; more formally, that the (approximate) truth of our current scientific theories provides at least the best and perhaps the only explanation for the empirical successes those theories enjoy.

Such realism stands challenged most powerfully by the *pessimistic induction* over the history of scientific inquiry (PI): thinkers ranging from Pierre Duhem and Henri Poincaré to Larry Laudan have repeatedly pointed out that the history of science exhibits a parade of eminently successful theories that have nonetheless been subsequently rejected. And they ask why we should not believe that the same fate awaits our own successful theories.[1]

* My thanks to David Malament, Jarrett Leplin, Philip Kitcher, Stathis Psillos, Larry Laudan, Jeff Barrett, Pen Maddy, Anjan Chakravartty, Alan Nelson, Alexander Rosenberg, Arthur Fine, and the members of a number of graduate seminars in which I have addressed these issues for helpful discussion, suggestions, and comments. This material is based upon work supported by the National Science Foundation under Grant No. SES-0094001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation (NSF).

[1] Two anonymous referees rightly note that I here follow the regrettably common practice of characterizing the relevant sorts of success at a painfully vague or intuitive level, though I am happy to accept the appealingly commonsensical account of such success (grounded in the potential for solving practical problems of prediction and intervention) offered by Kitcher (in "Real Realism: The Galilean Strategy," *The Philosophical Review*, CX, 2 (April 2001). 151–97, especially pp. 166–67). As a general

0022-362X/03/0011/553–72 © 2003 The Journal of Philosophy, Inc.

Of course such a challenge cannot depend upon agreement with current theories as a test of truth, so historical examples offer classically inductive evidence of the *falsity* of our own successful theories only if we have some other ground for regarding these successful predecessors as definitively refuted. But the challenge need not be formulated in this way in any case. The ghostly historical procession shows that past theories have been quite successful in just the ways that impress contemporary realists while nonetheless making fundamental claims about nature *inconsistent* with those of present successful theories (as well as those of other past successful theories in the same domains). Since *at most* one theory in any such inconsistent set can be true, we are provided with abundant inductive evidence that empirical success cannot possibly be a reliable indicator of the truth about nature, no matter what the truth is: any general practice of inferring from such successes to the approximate truth and/or reference of the theories that enjoy them would routinely if not invariably have to lead us astray. Alternatively, we can see the realist inference from success to approximate truth and/or reference as *self-undermining*, for if the success of current theories leads us to conclude that they are approximately true and/or referential, this implies in turn that many past theories must have been radically false and/or nonreferential *despite* being successful, undermining our original ground for concluding that current theories are approximately true and/or referential in the first place.[2]

Recent decades have witnessed not only impressive realist attempts to block or blunt this historical challenge by engaging the details of the history of science itself, but also the first serious efforts to recruit those details to the realist cause. While this engagement is certainly welcome, I will here argue that the most promising and influential of these efforts (including that of Clyde Hardin and Alexander Rosenberg, and, more recently, those of Philip Kitcher, Stathis Psillos, Jarrett Leplin, and John Worrall) manage to achieve only Pyrrhic victories

---

matter, however, defenders of the PI have resisted tying themselves too closely to *any* particular conception of scientific success, seeking instead to argue that there is (at most) a difference of degree and not in kind between present and past theories with respect to *whatever* sorts of success realists suppose could only be explained by the truth of the theories which enjoy them (see, for example, Laudan's "A Confutation of Convergent Realism," reprinted in David Papineau, ed., *Philosophy of Science* (New York: Oxford, 1996), pp. 107–38).

[2] This last formulation is similar to that offered by John Worrall in "How to Remain (Reasonably) Optimistic: Scientific Realism and the 'Luminiferous Ether'," in D. Hull, M. Forbes, and R. M. Burian, eds., *P S A. 1994*, Volume 1 (East Lansing, MI: Philosophy of Science Association, 1994), pp. 334–42, here p. 334.

for realism—that is, "defenses" of the approximate truth of past theories and/or the referential status of their central theoretical terms that are forced to concede to the realist's opponent either just the substantive points that were in dispute between them or everything she needs for a convincing historical case against realism. Although the resulting positions might fairly be said to be realist in name only, the point is far from merely terminological: in each of the cases I will discuss, those who seek to use the historical record to defend realism seem unaware that they have sacrificed the substantive tenets of the realist position on the altar of its name.[3]

I. REFERENCE WITHOUT DESCRIPTIVE ACCURACY

One strategy of realist reply to the historical challenge suggests that Laudan's classic defense of the PI exaggerates the extent to which the central terms of the rejected past theories he considers should be judged nonreferential by present lights. Laudan seems to require, for example, that the descriptive claims theories make about their central posits be largely accurate in order for the central terms in those theories to refer successfully, and some critics have suggested that causal accounts of reference reveal this requirement to be prejudicially restrictive: Hardin and Rosenberg,[4] for instance, argue that because "[o]ne permissible strategy of realists is to let reference follow causal role," realists are free to regard the central terms of even those theories they regard as radically mistaken to have been referential

[3] Of course, even if I am right to think that recent realist efforts achieve only Pyrrhic victories for the realist cause, we will still be left with a difficult question about just how to regulate our beliefs about the world in response to our historical predicament. Although I have characterized the PI as a challenge to our scientific theorizing about "otherwise inaccessible domains of nature," the distinction between observables and unobservables which figures so centrally in other influential challenges to realism (including, most famously, Bas van Fraassen's in *The Scientific Image* (New York: Oxford, 1980)) does not seem especially apposite here, for our characterizations of many observable entities (for example, chemical elements, evolutionary adaptations, supernovae) would seem to be routinely grounded in just those sorts of theoretical conceptions of the natural world that stand challenged by the PI itself. In forthcoming work I argue that the historical challenge is correctly focused not on our beliefs about entities of a particular sort (for example, "unobservables") but instead on beliefs *arrived at* in a particular way: more specifically, beliefs about nature that are reached *eliminatively*, that is, by coming to embrace one belief among a set of possibilities by eliminating its competitors, whether we so reason about observable or unobservable aspects of the natural world. But we would do well to settle first whether *any* serious challenge survives the most sophisticated and dogged recent efforts to convince us that the history of science actually poses no significant threat to scientific realism at all, and I will therefore defer to another occasion the general question of precisely which sorts of contemporary scientific beliefs the challenge of history calls into question.

[4] "In Defense of Convergent Realism," *Philosophy of Science*, XLIX (1982): 604–15

after all. For example, because we regard the electromagnetic field as playing the causal role attributed to the ether by earlier physical theories, the realist may hold, they suggest, "that 'ether' referred to the electromagnetic field all along" (*ibid.*, pp. 613–14).

As a response to the PI, however, this misses the forest for the trees. The sort of account envisioned by Hardin and Rosenberg secures a history of successful reference for terms in discarded theories only by divorcing their reference from the question of the accuracy of those theories and thus *abandoning* the specifically theoretical beliefs of the very sort for which the realist hopes to convince us to share her realism in the case of current theories. But this runs the realist afoul of what we might call the "trust" argument: after all, if the central terms in past theories are held to be referential *despite* the fact that the theories in which they are embedded repeatedly turn out to be radically misguided, then the historical record seems to entitle the antirealist to her claim that we would be foolish to *trust* either the theoretical accounts of inaccessible domains of nature offered by successful contemporary scientific theories or the descriptions associated with their central theoretical terms. And trusting the accounts of such domains and entities given by (some) current theories is just what the realist hoped to convince us to do!

Precisely because this shortcoming of Hardin and Rosenberg's account is so easily recognized,[5] it is surprising that a version of the same problem confronts more recently influential and sophisticated realist approaches to the reference of theoretical terms. Kitcher,[6] for instance, argues convincingly that the particular *tokens* of a speaker's use of a natural kind term must be separated and her dominant linguistic intentions considered in assigning reference to them: thus the reference of some of Priestly's tokens of 'dephlogisticated air' are fixed by his intention to refer to the substance emitted in combustion, while the reference of other tokens is fixed by his intention to refer to the substance whose inhalation was rendering his breathing particularly light and easy or the substance he "exploded together" with "inflammable air" to produce water or nitric acid. In the former cases, Priestly's tokens of 'dephlogisticated air' fail to refer (as there is no substance emitted in combustion in the way Priestly imagines), but the latter tokens instead refer to oxygen. Likewise, some of Fresnel's

[5] See Worrall, "Structural Realism: The Best of Both Worlds?" *Dialectica*, XLIII (1989): 99–124, especially pp. 116–17; Laudan, "Discussion: Realism without the Real," *Philosophy of Science*, LI (1984): 156–62, especially p. 161; and Psillos, *Scientific Realism: How Science Tracks Truth* (New York: Routledge, 1999), especially chapter 12.

[6] *The Advancement of Science* (New York: Oxford, 1993), chapter 4.

tokens of 'light wave' fail to refer because their reference is fixed by their theoretical description as oscillations of molecules of the ether, while the reference of others is fixed by Fresnel's dominant intention to talk about light, however it is in fact constituted, and therefore refer to electromagnetic waves of high frequency.[7] Kitcher concludes that claims of referential failure for past theories are overstated, and that the heterogeneity of the different modes of reference-fixing in different contexts of use permitted past theorists to make many referential and indeed true claims about the world.

But for all its success in rescuing the terms in discarded theories from blanket assertions of referential failure, Kitcher's approach runs afoul of the trust argument in just the same way that Hardin and Rosenberg's much simpler appeal to pure causal theories of reference did: Kitcher manages to rescue the reference of the central terms in discarded theories only *on those occasions* in which the user's dominant linguistic intentions explicitly eschew those specifically *theoretical* descriptions (like 'the substance emitted in combustion' or 'the oscillations of molecules of the ether') associated with her terms. But surely it offers little comfort to the realist if we insist that some tokens of terms like 'dephlogisticated air', and 'light wave' in rejected theories referred after all while admitting that the relevant theoretical accounts and descriptions of those entities were mistaken about virtually everything except the fact that the entities in question played some causal role in producing observable phenomena, for it is (once again) ultimately those very theoretical accounts and descriptions which the realist hopes to defend in the case of current theories. Thus, Kitcher offers a welcome sophistication to our account of reference, but one that makes no progress whatsoever in defending realism from the historical challenge: his account shows how tokens of the central terms of past theories were referential (and past theorists were able to enunciate important truths) just where their being so (or doing so) did not depend upon those theories actually getting anything much right about the natural world.[8]

[7] This strategy of analysis is further developed in connection with a general causal-descriptivist account of reference for natural kind terms in Stanford and Kitcher, "Refining the Causal Theory of Reference for Natural Kind Terms," *Philosophical Studies*, XCVII (2000): 99–129.

[8] Of course, Kitcher's further argument (*Advancement of Science*, chapter 5) that those *parts* of superceded theories enabling them to be successful were also true (see also Psillos) would seem to require that at least some of the central terms of past theories referred successfully even on occasions when their reference was fixed in what he calls the "descriptive mode" *and the descriptions in question were theoretical* (and conversely, that where uses of past terms were sufficiently infected with false theory as to be nonreferential, they were not involved in or responsible for those theories' successes). I argue below that this argumentative strategy provides the realist's oppo-

By contrast, Psillos seems painfully aware that realists undermine their own cause in establishing mere referential continuity without descriptive accuracy for past theories. His own defense of realist reference (*op. cit.*, chapter 12) is carefully constructed to avoid this particular pitfall by requiring at least some of the descriptive information associated with terms by our theories to be accurate in order for those terms to refer. It is especially revealing, then, that his account nonetheless concedes to the realist's opponent everything she needs to make her historical case. Let us see why.

Psillos argues that because pure causal theories make referential continuity too easy to get (that is, terms cannot fail to refer), any convincing account of reference along these lines will have to be "causal-descriptivist"; that is, it will have to require not simply that a term refer to whatever causes the phenomena that occasion its introduction into the language, but in addition that some of the fundamental descriptions associated with the term actually be satisfied by the entity to which it refers. But not all the associated descriptions are equally important: he argues that "some descriptions associated with a term are less fundamental in view of the fact that the posited entity would play its intended causal role even if they were not true" (*op. cit.*, p. 297). The descriptions that really count, then, and which must be satisfied by an entity in order for a term to refer to it, are those making up what he calls the theory's "core causal description" of the entity in question: the descriptions that would have to be true in order for the entity to play the causal role the theory assigns to it. Thus, he argues, the term 'phlogiston' failed to refer because "there is nothing which fits a description which assigns to phlogiston the properties it requires in order to play its intended causal role in combustion" (*op. cit.*, p. 291), in particular "the property that it is released during the process of combustion" (*op. cit.*, p. 298).

It is crucial to Psillos's account, however, that the term 'ether' in nineteenth-century wave theories of light and electromagnetism turn out to be referential, for (in contrast to Kitcher) he acknowledges that the postulation of the ether (as a "dynamical structure" serving as a carrier for light waves) played a crucial role in the successes of those theories. He argues that this demand creates no problem for the realist, however, because our own term 'electromagnetic field' shares the core causal description associated with the 'ether' of nineteenth-century optics, and so, he claims, the latter term referred to

---

nent with all the resources she needs to make a convincing historical case against realism.

the electromagnetic field itself.[9] Although current physical theories regard the beliefs of nineteenth-century optical and electromagnetic theorists about the *nature and constitution* of the ether to have been quite radically and fundamentally misguided, Psillos insists that such beliefs did not form any part of the actual core causal description of the term 'ether'. That is, the false beliefs of nineteenth-century theorists about the material and mechanical character of the ether were not part of the description that an entity would have to satisfy in order to play the causal role assigned to the ether by nineteenth-century wave theories of light and electromagnetism.

Of course, this account of the matter invites the realist to choose the core causal descriptions she associates with the central terms of past theories rather carefully, with one eye on current theories' claims about nature, so there is more than a whiff of ad hoc-ery about the proposal. But even if we set this worry aside and assume that the realist can delicately titrate the core causal descriptions she associates with the crucial terms in successful past theories so as to render them referential by the lights of current theories, Psillos's victory will *nonetheless* remain a Pyrrhic one. The reason is that this case for the referential status of central terms in successful past theories simply invites from the historical record a renewed form of the pessimistic induction itself, this time concerning our ability to distinguish (at the time a theory is a going concern) which of our beliefs about an entity are *actually* part of its core causal description. To see why, let us look more closely at the one case of reference Psillos examines in detail: the luminiferous ether of nineteenth-century optics.

The core of the problem is that nineteenth-century theorists themselves strenuously disagreed with *the very assessment Psillos offers* of what would have to be true of an entity in order for it to play the ether's causal role, that is, with his very claim about *which* descriptions of the ether enter into its core causal description. They considered whether the ether could play the role ascribed to it in propagating light and other electromagnetic waves without consisting of a mechanical medium of some kind *and explicitly denied that it could*! Maxwell himself concludes *A Treatise on Electricity and Magnetism*[10] with the following

[9] Actually, Psillos's claim seems better suited to the early twentieth-century conception of the electromagnetic field than the contemporary one: it is at least contentious to describe the electromagnetic field recognized by contemporary quantum electrodynamics as one in which light waves *propagate* at all (c.f. Psillos, p. 296). But I will not rely on this point in what follows: instead I will show that Psillos's realist victory is hollow even if we treat this early twentieth-century conception of the field as our own.

[10] 3rd edition in two volumes (London: Oxford, 1955 [1873]).

resolute insistence that the ether must be a material medium *in order to play the causal role* of propagating energy waves:

> If something is transmitted from one particle to another at a distance, what is the condition after it has left the one particle and before it has reached the other? If this something is the potential energy of two particles, as in Neumann's theory, how are we to conceive this energy as existing in a point of space, coinciding neither with the one particle nor with the other? *In fact, whenever energy is transmitted from one body to another in time, there must be a medium or substance in which the energy exists after it leaves one body and before it reaches the other,* for energy, as Torricelli remarked, 'is a quintessence of so subtile a nature that *it cannot be contained in any vessel except the inmost substance of material things*'. Hence all these theories lead to the conception of a medium in which propagation takes place, and if we admit this medium as an hypothesis, I think it ought to occupy a prominent place in our investigations, and that we ought to endeavor to construct a mental representation of its action, and this has been my constant aim in this treatise (*ibid.*, Volume II, p. 493, my emphasis).

Maxwell's positive conception of the field is a matter of some legitimate controversy, but here he quite explicitly denies that anything *besides* a mechanical or material substance is capable of playing the causal role he assigns to the ether: he insists that to avoid the patent (or at least inconceivable) absurdity of "energy...existing in a point of space" we must recognize the existence of a material medium filling the space between bodies, for energy "cannot be contained in any vessel except the inmost substance of material things." Thus, Maxwell self-consciously insisted that the "core causal description" of the ether *included* the very beliefs about the ether's material and mechanical character that Psillos grants are not satisfied by the modern conception of the electromagnetic field as a distinct entity ontologically on par with matter itself.[11]

[11] Of course, the contemporary conception of the electromagnetic field ascribes to it a well-defined (local) mass-energy content and it may therefore be said to qualify as a "material substance" in *some* current sense of this term. But the sense in question is far removed from the requirements Maxwell had in mind in insisting that the ether must be a "medium or substance" or denying that energy can reside anywhere "except the inmost substance of material things" (and Psillos seems to grant as much in recognizing that nineteenth-century theorists' beliefs about the nature and constitution of the ether have turned out to be mistaken). Thus, we must understand Maxwell as insisting that nothing besides a material or mechanical substance as *he* understood these notions could play the causal role he assigned to the ether, and it is the inaccuracy of this judgment (by modern lights) which suggests a renewed PI concerning our ability to discern just which parts of the descriptions we associate with terms in our successful theories are really part of their "core causal descriptions." My thanks to David Malament for helpful discussion on this point.

Thus, even if Psillos can convincingly argue that the *actual* "core causal description" of the ether does not include claims about its material or mechanical character, he will be forced to concede that the carefully considered judgments of leading scientific defenders of the theory concerning which of the descriptions associated with its central terms must be satisfied by an entity in order for it to play the causal role associated with the term (that is, which features figure in the actual core causal description) have proved to be unreliable. What this suggests, of course, is that we cannot rely on our *own* judgments about which of the descriptions we associate with our own terms are *genuinely* part of their own core causal descriptions, delivering us back into the arms of the trust argument with respect to current theories.[12]

Once this underlying problem with Psillos's strategy is recognized, further historical examples are relatively easy to identify. In his celebrated defense of the "continuity of the germ-plasm," August Weismann[13] argues vigorously that the germinal or hereditary materials are located in the nuclei of cells and that they must have an internal structure in which different parts of this germ-plasm are responsible for the production of varying characteristics in different cells. Further, he argues explicitly that the germinal materials must be separated and reduced in each subsequent cell division, until each cell contains only the tiny amount necessary for determining its own characteristics. He argues that there simply is no other way to account for the fact that different parts of the body possess different characteristics, and, in particular, insists that the cells of the body could not possibly be heterogeneous if each cell carried precisely the same hereditary materials:

> As the thousands of cells which constitute an organism possess very different properties, *the chromatin* which controls them *cannot be uniform; it must be different in each kind of cell.*
>
> The chromatin, moreover, cannot *become* different in the cells of the fully formed organism; the differences in the chromatin controlling the cells must begin with the development of the egg-cell, and must increase

[12] Psillos admits that there is an element of rational reconstruction in settling on the appropriate core causal descriptions for terms in past theories, but this glides over the real problem we have noted: without justifiable confidence in our ability to specify accurately core causal descriptions for theoretical entities, we do not know which features of our *own* theories are rightly included in the core causal descriptions associated with *their* central terms, and thus have no way to pick out which of our own actual beliefs we can trust.

[13] *The Germ-Plasm: A Theory of Heredity*, W. Newton Parker and Harriet Rönnfeldt, trans. (New York: Scribner's, 1893 [1892])

> as development proceeds; for otherwise the different products of the division of the ovum could not give rise to entirely different hereditary tendencies. This is, however, the case. Even the first two daughter-cells which result from the division of the egg-cell give rise in many animals to totally different parts.... The conclusion is inevitable that the chromatin determining these hereditary tendencies is different in the daughter-cells (*ibid.*, p. 32, original emphasis).

The point here is not simply that Weismann made (by current lights) some mistakes about the constitution of the germ-plasm, still less is it important whether these false beliefs prevented particular tokens of 'chromatin' or 'germ-plasm' from referring on particular occasions of use. Rather, the point is that Weismann argued quite explicitly that *no entity could possibly play the causal role that he assigned to the hereditary material without being separated and parceled out differently to different cells*, offering further evidence of our historical unreliability in judging which theoretical descriptions are genuinely part of a term's core causal description.

Psillos's strategy is similarly undermined by cases in which the term in question is presently judged to be uncontroversially *non*referential. Consider Antoine Lavoisier's claim, defending the caloric fluid theory of heat in his 1785 *Memoir on Phlogiston*, that

> One can hardly think about these [thermal] phenomena without admitting the existence of a special fluid [whose accumulation causes heat and whose absence causes cold]. It is no doubt this fluid which gets between the particles of bodies, separates them, and occupies the spaces between them. Like a great many physicists I call this fluid, whatever it is, *the igneous fluid, the matter of heat and fire*.[14]

The point is not, of course, that Lavoisier was wrong about the need to posit caloric fluid. Rather, Lavoisier here suggests that nothing *besides* a subtle fluid could play the role in causing thermal phenomena that he assigns to caloric (but that contemporary theorists assign to molecular motion). This assessment is drawn even more explicitly in terms of caloric fluid's causal role in Lavoisier's later work:

> It is difficult to conceive of these phenomena without admitting that they are the result of a real, material substance, of a very subtile fluid, that insinuates itself throughout the molecules of all bodies and pushes them apart.... This substance, whatever it is, is the cause of heat, or in

[14] "Mémoire sur phlogiston," as translated in A. Donovan, *Antoine Lavoisier: Science, Administration, and Revolution* (Cambridge: Blackwell, 1993), here p. 171, original emphasis, translation modified.

other words, the sensation that we call heat is the effect of the accumulation of this substance....[15]

Lavoisier's nonreferential account of caloric shows, no less than Maxwell's supposedly referential discussion of ether and Weismann's putatively referential account of the germ-plasm or chromatin, that we have historically been unreliable (again, by present lights) in judging which descriptions must be satisfied by an entity or what characteristics it must have in order for it to play the causal role assigned to it by a particular theory. Like Kitcher's then, Psillos's defense of the referential status of some central terms in past theories comes only at the cost of leaving us unable to be confident that any particular descriptions we associate with a referring term (even ones we presently regard as absolutely central and/or indispensable to fulfilling its causal role) will be retained in the development of further theoretical science. He therefore achieves a victory for realism only at a price that realists cannot afford to pay.

II. DILUTING APPROXIMATE TRUTH

Related difficulties afflict the complementary strategy of suggesting that Laudan is too quick to deny that his examples of superseded theories were approximately true. There is a further complication here, in that Laudan takes advantage of an argumentative shortcut in making his case against the approximate truth of rejected theories: he assumes explicitly that the failure of a theory's central terms to refer (in a sense that requires descriptive accuracy) ensures that the theory is not approximately true, arguing that "the *realist would never want to say that a theory was approximately true if its central theoretical terms failed to refer.* If there were nothing like genes, then a genetic theory, no matter how well confirmed it was, would not be approximately true."[16] Hardin and Rosenberg (*op. cit.*) protest, insisting that a case

[15] *Traité de chimie*, in Volume 1 of *Oeuvres de Lavoisier*, 6 volumes, volumes 1–4, J. B. Dumas, ed. volumes 5–6, Edouard Grimaux, ed. (New York: Johnson Reprint Corporation, 1965 [originally published in Paris. Imprimerie Impériale, 1862–1893]) p. 19, my translation

[16] "A Confutation of Convergent Realism," p. 33, original emphasis. Laudan avails himself of this shortcut in part because he argues that realists have failed to provide any account of approximate truth on which the presumption that the approximate truth of a theory implies or entails its likely success can be defended. He suggests that typically this presumption is uncritically assumed to follow from the unobjectionable fact that a perfectly true (and, we might add, complete) theory would be perfectly successful—an inference that he insists is patently invalid—and he challenges realists to provide an analysis of approximate truth on which the presumption of success for approximately true theories is defensible. Thus, Laudan cannot afford to pin his argument on any particular conception of approximate truth itself.

like classical Mendelian genetics illustrates how a theory might have central terms that fail to refer by Laudan's descriptive criteria (for they insist, perhaps implausibly, that contemporary genetics recognizes nothing like a Mendelian gene) but nonetheless earn a judgment of approximate truth by current practitioners. Thus, they claim, Laudan's case relies inappropriately on the reference of central terms as a minimal condition for approximate truth.[17]

Laudan's own response to this objection asks by what right Hardin and Rosenberg "take contemporary theories as benchmarks of what there is and how it behaves" once they have granted that a successful theory (like Mendel's) may be very wide of the ontological mark and thereby undermined the realist's abductive argument itself.[18] But this reply risks missing the point of the objection: Hardin and Rosenberg can respond that they are simply showing how a current theory, *if* true, could ground the judgment that a particular past theory was both nonreferential and approximately true, thereby invalidating Laudan's argumentative shortcut from failures of reference to failures of approximate truth, and thus undermining the case he makes against the realist's abductive inference in the first place.

More effective, therefore, is to ask what weight this sort of objection to Laudan's argumentative shortcut is ultimately supposed to carry. Even if we grant Hardin and Rosenberg that the relationship between Mendelian and contemporary genetics illustrates how one theory can be (by another's lights) both nonreferential and approximately true, *this is surely not what the verdict of our own actual contemporary theories would be concerning most, if not all, of the other theories on Laudan's historical hit parade.* If, as the realist would have it, our contemporary theories are true, then we are inclined to insist that the many eminently successful theories included on Laudan's list (including the phlogiston theory of chemistry, the caloric theory of heat, the vital forces theories of physiology and embryology, and various optical and electromagnetic ether theories) are *not* in fact true, not even approximately so: the relationship between current theories and those on Laudan's list is simply not, in general, the one that Hardin and Rosenberg claim obtains between Mendelian and contemporary molecular genetics. Furthermore, the claim that these successful past theories are not even approximately true by contemporary lights is indeed

[17] Not all of Laudan's examples proceed in this fashion, however: he takes explicit notice of several theories whose central terms *did* refer, but which he suggests were nonetheless not approximately true ("A Confutation of Convergent Realism," pp 123–24).

[18] "Realism without the Real," p. 159.

strongly *supported* by the fact that current theories hold that there is nothing like the central posits of those superseded theories, even if this fact is not alone sufficient to guarantee their radical falsity by contemporary lights.

Of course, Hardin and Rosenberg might instead have in mind some sense of "approximate truth" in which the literal truth of current theories is indeed consistent with the approximate truth of such classic success stories as the phlogiston theory of chemistry, the caloric theory of heat and the theories of the optical and electromagnetic ethers, but if so they will quickly find themselves back in the jaws of the trust argument. This is because the "approximate truth" of past successful theories, in *any* sense that is consistent with being as fundamentally and profoundly mistaken about the constitution of nature as these famous predecessors were by the lights of current theories, simply does not cut against the antirealist's insistence that the historical record shows why it would be a mistake to trust or believe the theoretical accounts of nature they offer or that doing so would have routinely led us badly astray in the past. Thus the realist will win a battle over something she calls "approximate truth," but again lose the war over realism, for her opponent's skepticism about the accounts of nature offered by current successful theories will rightfully survive her concession that those theories may well be "approximately true" in the attenuated sense the realist has managed to defend for past theories.

This same point undermines Kitcher's suggestive analogy between the response of modest scientific realists to the historical record and "that of the author who confesses in her preface that she is individually confident about each main thesis contained in her book but equally sure that there's a mistake somewhere."[19] The analogy is clever, for it invites us to see the realist's opponent as simply carping over our inability to attain an unreasonable standard of accuracy in what is admittedly a difficult business. Nonetheless, even when we restrict our attention (as Kitcher insists) to those historical cases of past theories grounding "predictions and interventions that were numerous, diverse, and hard to achieve,"[20] the analogy proves to be seriously misleading, for contemporary theories are surely not rightly thought to hold such successful predecessors as the wave theory of light, Newtonian mechanics, and the caloric theory of heat to have been mistaken simply in matters of minor detail comparable to a misplaced

[19] "Real Realism. The Galilean Strategy," p. 171, see also Kitcher's *Science, Truth, and Democracy* (New York. Oxford, 2001), pp. 18–19.

[20] *Science, Truth, and Democracy*, p. 19.

footnote, a speculative musing, or even an overstated conclusion. Instead, they hold these illustrious predecessors to have been deeply and thoroughly mistaken in their most central claims about the constitution and/or operation of nature, a situation more analogous to the author having been fundamentally mistaken in the principal thesis of her book or the central contentions she was concerned to advance.

Other recently influential efforts to defend realism from the historical record turn out to dilute approximate truth in ways that are similarly self-defeating. Leplin's extended defense of realism,[21] for example, champions the inference from a theory's success in making *novel* predictions (in a precisely specified sense of novelty) to what he calls its "partial truth" (*ibid.*, p. 127). Leplin modestly aims to defend only what he calls minimal epistemic realism: the claim that there *are* epistemic conditions that would warrant a realist attitude towards a theory, not that any present theory actually satisfies such conditions. But the partial truth Leplin argues we can infer from a theory's novel predictive success is nonetheless too meager to render the prospect of this inferential entitlement of any real significance in the debate over realism.

This is because Leplin rightly sees that he cannot infer any particular degree, kind, or respect of partial truth (in the sense of representational accuracy[22]) from any particular degree, kind or respect of success in novel prediction. He grants that he is "vague by default as to how much novel success merits what level of confidence in representational success," explicitly denies that "the amount of novel success provides a measure of the degree of representational accuracy achieved," and acknowledges that novel predictive success does not warrant attributing "a particularly high level of accuracy, because we have no way to determine what forms of inaccuracy might be irrelevant to the observable situation" (*ibid.*, p. 128). But with no inferential connection between degrees or kinds of novel predictive success and degrees or kinds of representational accuracy, we will never be able

[21] *A Novel Defense of Scientific Realism* (New York: Oxford, 1997)

[22] Leplin articulates two quite distinct senses of partial truth. The "pragmatic" sense seeks to capture what practicing scientists commit themselves to in accepting theories, both prospectively (contrasting the partial truth of present theories with falsity) and retrospectively (contrasting the partial truth of past theories with unqualified truth), while the "metaphysical" sense of partial truth is understood in terms of representational accuracy and is what we are supposedly entitled to infer from a theory's success in novel prediction. It is only this second, metaphysical sense of partial truth that will concern us here. My thanks to Leplin for clarifying this and other aspects of his work in correspondence, though he would not be at all satisfied by the conclusions I reach below.

to say anything more about any theory (even a merely possible future theory) than that it enjoys "some" degree of representational accuracy, no matter how much novel predictive success it has enjoyed. And this (unimprovable) claim is trivially satisfied: after all, Aristotelian mechanics, creationist biology, caloric thermodynamics, and phlogistic chemistry all enjoy "some" degree of representational accuracy, too—none of these theories is wrong or misleading about absolutely everything, not even everything fundamental.[23]

Thus, although he intends to eschew the traditional realist commitments he regards as indefensible, Leplin cannot, I suggest, defend even a minimal realism with such a feeble connection between novel predictive success and representational accuracy or partial truth. Instead his account dilutes the notion of partial truth to the point that the historical record of such partially true theories supports rather than opposes the antirealist's skepticism about the descriptive or representational accuracy of current (and even future) successful theories.[24]

Perhaps for this reason, Leplin sometimes seems tempted to assert a more fine-grained connection between novel predictive success and a particular degree or kind of representational accuracy: responding to the PI, he suggests that the fundamental theoretical mechanisms employed by theories that enjoyed novel predictive success have not been overturned by subsequent developments (*op. cit.*, p. 145). But he is fully aware of the problem that such a suggestion creates: the classic leading example of novel predictive success—the prediction of the Poisson bright spot—was made by a theory now regarded as

[23] For example, while none of creationism's causal or explanatory mechanisms are accepted in current biology, its theoretically motivated division of organisms into species is nearly identical to the leading approach in contemporary evolutionary theory (the biological species concept). The other three examples arguably enjoy some degree of representational accuracy by present lights even at the level of causal and explanatory mechanisms: for example, one of the important respects in which Kuhn famously suggests (in *The Structure of Scientific Revolutions*, 3rd edition (Chicago: University Press, 1996), pp. 206–07) that "Einstein's general theory of relativity is closer to Aristotle's [mechanics] than either of them is to Newton's" is presumably that general relativity (like Aristotelian mechanics) recognizes gravitational motion as itself a "natural" state of motion (that is, along a "straight" trajectory in curved spacetime) not requiring further causal explanation rather than a deflection from natural (inertial) motion as in Newtonian mechanics.

[24] This also seems the appropriate response to Larry Sklar's contention (in *Theory and Truth* (New York: Oxford, 2000) especially section 4.1) that our best current theories should be viewed in light of history as "on the road to truth," "pointing towards the truth," or "heading in the right direction", however, it is far from obvious that Sklar would disagree with this assessment and his central contention is that the most interesting and important issues are simply obscured at this level of abstraction in any case.

radically misguided, namely, Fresnel's formulation of the wave theory of light, with its conception of light waves as oscillations of the molecules of a material ether. Therefore, Leplin's "direct" response to the skeptical induction (*op. cit.*, p. 146f.) goes on to appeal hopefully to Kitcher's suggestion in *The Advancement of Science* that those *parts* or *aspects* of rejected theories actually responsible for their successes (in this case, particularly their successes in novel prediction) have been preserved in or ratified by subsequent theorizing about nature. That is, Kitcher argues (as does Psillos, *op. cit.*) that past practitioners had *selective* confirmation for only those *parts* of past theories that have turned out to be *true*.[25]

This same "selective confirmation" approach also proves central to Kitcher's more recent defense of realism using what he calls the "Galilean Strategy" of generalizing the inference from success-to-truth in everyday contexts like card games and attempts to use the subway system (where its reliability can be checked) to that of theorizing about the natural world.[26] Kitcher's defense of this generalization depends, as he recognizes, on establishing convincingly that theories' "past successes stem from parts of the theories that are approximately correct,"[27] that is, on the ability of the strategy of selective confirmation to turn apparently failed past instances of the success-to-truth inference strategy into successful ones; otherwise we have compelling reasons to doubt that the reliability of the everyday success-to-truth inference survives its Galilean importation into the quite different context of scientific theorizing.

Of course, to evade the trust argument, the appeal to selective confirmation will have to provide what we might call "prospectively applicable" criteria of selective confirmation: that is, criteria that could have been applied to past theories at the time and can now be applied to our own theories *in advance of any future developments* to say just which parts of past theories were (and just which parts of present theories are) genuinely confirmed by the successes they enjoy. Else-

[25] In private correspondence, Leplin has indicated that he is disinclined "to try to turn what is clearly a problem for realism into a positive argument" in the way I suggest here, because even when the theoretical mechanisms responsible for novel predictive success survive in subsequent theories they are sometimes "so radically reconceived...that I do not hold much hope for founding upon their retention greater specificity as to [what] descriptive content novel success warrants a commitment to." He thus rejects the response I explore here and will have to accept instead that success in novel prediction warrants only an inference to "some" degree of representational accuracy, an inference which I have argued trivializes the realism it seeks to defend.

[26] See "Real Realism: The Galilean Strategy," and *Science, Truth, and Democracy*

[27] "Real Realism: The Galilean Strategy," p. 170.

where,[28] I have argued in detail that the actual criteria of selective confirmation on offer fail this test: the grounds on which Kitcher proposes to distinguish "working" from "idle" posits fail to discriminate the rejected posits of past theories (like 'ether') from the posits he hopes to defend in present ones (like 'genes'), while Psillos's claim that scientists' own judgments of selective confirmation have proved to be reliable itself depends upon a highly selective reading of the historical evidence.

We need not review the details of these arguments, however, to see why this strategy of selective confirmation offers us yet another instance in which the realist reply concedes everything her opponent needs to make her historical case, for a central aspect of the problem can be extracted just from the historical sources noted above. The passages cited from Maxwell, Weismann, and Lavoisier illustrate that we have repeatedly misidentified those parts, features, or aspects of our theories that (by realist lights) were genuinely implicated in or required for their successes: as in the case of Psillos's core causal descriptions, then, we here have the materials for a renewed version of the pessimistic induction itself, this time concerning our ability to determine, at the time a theory is a going concern, which parts, features, or aspects are *actually required for the successes* of that theory. And as the case of Maxwell illustrates in striking detail, our ability to draw this distinction accurately is suspect even in the case of theories' successes in making novel predictions: Maxwell was as clear and explicit as he could be in insisting that the very *coherence* of the wave theory's successful application required the existence of a material, mechanical, substantival ether. Without confidence in our ability to pick out the parts of theories needed for their successes while those theories are live contenders, however, we are in no position to identify those parts or features of our own theories we may safely regard as accurate descriptions of the natural world (even though we know that not all are) and the realist's opponent again seems entitled to the conclusion she wanted all along.

A version of this same problem threatens to undermine the contention of Worrall's "Structural Realism"[29] that *structural* commitments typically survive the demise of the predictively successful theories in which they occur (by contrast with the theory's "content,"[30] "ontol-

[28] "Selective Confirmation: No Refuge for Realism," forthcoming in *Philosophy of Science*, supplemental volume to *P.S.A. 2002*.

[29] Developed most fully in "How to Remain (Reasonably) Optimistic: Scientific Realism and the 'Luminiferous Ether'," and "Structural Realism: The Best of Both Worlds?"

[30] "Structural Realism," p. 117; "How to Remain (Reasonably) Optimistic," p. 340.

ogy,"[31] or its claims about the "nature" of entities[32]) and are therefore the legitimate objects of justified realist confidence. It is not at all clear that we can plausibly distinguish the merely "structural" claims of even current theories from their "content" or claims about the "nature" of entities, not to mention assess the record of past practitioners in making such discriminations successfully, but (to invoke the now familiar refrain) unless we can do both of these things we will not know what aspects of current theories are the ones we can safely regard as accurately describing features of the natural world. Appeals to vague intuitions simply will not do here: at best such an intuitive criterion renders the problem of reliable application especially acute, forcing the structural realist to qualify her beliefs in even the structural claims of a theory by her level of confidence that they are indeed structural claims. But even worse, a merely intuitive criterion of structure seems to run afoul of the historical record in even central cases: What prevents, for example, Weismann's insistence that germinal materials must be parceled out differently to heterogeneous parts of an organism's body from being a claim about the *structure* of inheritance and ontogeny?

Perhaps in light of such difficulties, Worrall sometimes suggests that the structural commitments of a theory consist simply of its equations or the abstract mathematical relationships it posits.[33] But there is something extremely misleading in saying even that the abstract mathematical relationships posited by past successful theories have described the "structure" of the natural world in ways that are still embraced by current theories. Francis Galton's ancestral law of inheritance, for instance, was the central mathematical formalism and the most predictively successful aspect (see below) of his "stirp" theory of inheritance: it claimed that the germinal materials received by each individual (and determining her characteristics) are composed of fractional contributions made directly from each ancestor, in the proportion

$$1/4p + 1/8pp + 1/16ppp...$$

where p is the contribution from the parental generation (on each side), pp from the grandparental, and so on.[34] Of course it is true

[31] "How to Remain (Reasonably) Optimistic," pp. 336, 341.

[32] "Structural Realism," pp. 117–18; "How to Remain (Reasonably) Optimistic," p. 334.

[33] "Structural Realism," pp. 118–20, "How to Remain (Reasonably) Optimistic," p. 340.

[34] Galton's clearest formulation of the law of ancestral inheritance is found in his "The Average Contribution of Each Several Ancestor to the Total Heritage of the Offspring," *Proceedings of the Royal Society of London*, LXI (1897): 401–13.

enough to say that the fractional relationships described by Galton's ancestral law show up *somewhere* in the description of inheritance provided by contemporary genetics; consider the following passage from Robert Olby's seminal *Origins of Mendelism*:[35]

> Today Galton's ancestral law of inheritance still stands as a mathematical representation of the average distribution of continuously varying characters in a population of freely outbreeding individuals not subject to selection. It serves as a basis for predicting the average distribution of such characters in the population. It tells us that on an average a grandparental deviation will be diminished to one-eighth of its original magnitude in the grandchildren. The Mendelian theory, on the other hand, tells us that only one in eight grandchildren will have received his grandparent's genes for the said deviation. Expressed as averages for a population, however, both theories give the same prediction (*ibid.*, pp. 81–82).

But contemporary genetics does not recognize the fractional relationships expressed in Galton's law of ancestral heredity (that is, 1/4 from the parental generation (on each side), 1/8 from the grandparental, 1/16 from the great-grandparental, and so on, *ad infinitum*) as describing any fundamental or even especially significant aspect of the mathematical structure of inheritance.[36] By present lights, it would be extremely misleading, if not outright mistaken, to say that even the mathematical structure expressed by Galton's Ancestral Law is preserved in contemporary genetics. Thus, to evade the challenge of the historical record, the mathematized version of structural realism will have to retreat simply to the dogged insistence that such chunks of mathematical formalism as Avogadro's number, Fresnel's equations for the transmission of light, and Galton's Ancestral Law of Inheritance will be recoverable in *some* way, *some*where, *some*how from future science. This, of course, is a far cry from giving us *any* claim we can rely on as an accurate description of (even just the structure of) the natural world, and therefore invites the trust argument again with a vengeance. Like the others we have examined, then, Worrall's structuralist defense of a realist position seems to give the game away entirely: it either leaves us with no justifiable confidence in our ability to distinguish clearly those claims of current theories about the natural

[35] New York: Schocken, 1966.

[36] This seems especially clear if we keep in mind that Galton's law is expressed in terms of *generational* contributions to the stirp of the offspring from each side: it is, for instance, the grandparental *generation* on each side which contributes 1/8 to the makeup of the stirp as a whole, and thus each grandparent makes only 1/16 of the stirpal contribution.

world on which we may rely, or it forces us to draw such a distinction in a way that invites yet another renewed pessimistic induction over the historical record itself.

III. CONCLUSION

I have tried to identify a systematic pattern of argument that consistently reappears among recent serious and subtle efforts to recruit the historical record to the defense of scientific realism and to argue that this pattern of argument ultimately undermines the very cause it is intended to advance. Defenses of the reference of central terms in past theories like those offered by Hardin and Rosenberg and Kitcher simply give up the substantive tenets of realism they sought to defend, while Psillos's appeal to core causal descriptions, Leplin's appeal to partial truth, Kitcher's and Psillos's appeals to the selective preservation of those parts of our theories involved in their successes, and Worrall's appeal to historical continuity in structural commitments are all forced either to defend realist inferential entitlements that are so weak as to capitulate to the realist's opponent on the question of whether we can safely trust the accounts of nature given by current or future successful scientific theories, or to rely on some discriminatory ability with respect to features, aspects, or claims of current theories whose reliability is itself vulnerable to a compelling historical challenge. A convincing defense of realism from the specter of the historical record will have to do better.

P. KYLE STANFORD

University of California/Irvine

GOODMAN'S PROBLEM AND SCIENTIFIC METHODOLOGY*

Nelson Goodman's "new riddle of induction"[1] shows that there is more than one way to project from data, in a way consistent with traditional "formal" constraints on such inductions and projections. From a data set of emeralds in which all are found to be green, we can infer (fallibly) that all emeralds are green, but we can also apparently infer that all emeralds are *grue*, where an object is grue if and only if it has been observed before now and is green, or has never previously been observed and is blue. The riddle asks: What makes the "green" induction better than the "grue" one?

I. INTRODUCTION

Half a century after Goodman, a huge variety of solutions has been proposed. There is no consensus on which solution is best, but many of the solutions have a feature in common. They assume that the two inductions share all the features that might be addressed by a general discussion of good methods of projection *within* science. Here I have in mind the kind of discussion that might be found in a statistics textbook, or a book describing the methods of data analysis used in some particular scientific field.

The two rival inductions are usually presented very schematically, of course: all observed *F*s are *G*, so probably all *F*s are *G*. But it is often thought that no real difference is made if we fill in more detail, and make the cases more scientifically life-like. We might dress up the two inductions in the garb of estimations from samples using statistical techniques, but that will not affect the basic problem. We will still have a set of emerald observations that is equally compatible with a normal-looking "green" projection and with an abnormal-looking "grue" projection.

So in a sense, the philosophical literature has often assumed that the statistics and data analysis textbooks used in science really need a kind of "chapter zero." This chapter zero would lay down constraints

* Earlier versions of this material were presented to colloquia and conferences at Stanford University, Australian National University, University of California/Irvine, and University of California/Berkeley. I am especially grateful to Aldo Antonelli, Branden Fitelson, Brian Ellis, Ben Escoto, Alan Hayek, John MacFarlane, Simon May, Rajani Rajan, and Brian Skyrms for comments and correspondence.

[1] Goodman's most famous discussion of his problem is in his *Fact, Fiction, and Forecast* (Cambridge: Harvard, 1955). For a collection surveying solutions proposed in the last fifty years, see Douglas Stalker, ed., *Grue!* (Chicago: Open Court, 1994).

0022-362X/03/0011/573–90 © 2003 The Journal of Philosophy, Inc.

on inference that are in a sense prior to the constraints discussed in the book itself. These constraints would impose restrictions on which predicates belong in data analysis at all. Some predicates, like "grue," are just *nonprojectible*, and should be ruled out or at least downgraded ahead of time, as the basis for inferences from data. Suggested bases for this restriction vary widely (from the metaphysical to the conventional), but the basic idea is very common. Of course, philosophers assume that in practice the enculturation of a young scientist will informally impart most of the content of the absent chapter zero. There is usually no need to make chapter zero explicit, except when doing philosophy.[2]

I will argue against this approach to Goodman's problem. I will also, roughly speaking, propose a new solution, but this solution draws very heavily on the work of others. The solution is also old in another sense; it rests on an appeal to some standard ideas in scientific data analysis. Further, part of my "solution" consists in making a distinction between different aspects of the problem, which require different kinds of treatment. What we have come to call "Goodman's problem" is something of a mixture. Philosophical discussions of confirmation have assimilated issues and problems that have very different properties. This finding has general consequences for the philosophical project of giving a "theory of induction." These claims about the disunity of Goodman's problem will be introduced at the end of the article, however.

As a preliminary, I should also note that some Bayesian analyses of Goodman's problem are important exceptions to my generalization about how the literature has handled Goodman's problem. Here I have in mind some versions of the idea that the solution involves a difference between the *prior probabilities* of the two hypotheses about emerald color.[3] I will not argue directly against this approach, and will say little about Bayesianism in this article. However, the core of my proposals could be embedded within a Bayesian framework.

II. THE JPB PROPOSAL

This section and the next will develop a proposed solution to Goodman's problem. Although this solution is based on concepts used

[2] John Pollock has expressed the alleged need for this supplement to standard statistical theory explicitly: "It seems likely that...[a] strong projectibility constraint should be imposed upon all familiar patterns of statistical inference. The need for such a constraint seems to have been ignored in statistics"—"The Projectibility Constraint," in Stalker, ed., p. 141.

[3] See, for example, Colin Howson and Peter Urbach, *Scientific Reasoning: The Bayesian Approach* (Chicago: Open Court, 1989), chapter 4; and Elliott Sober, "No Model, No Inference: A Bayesian Primer on the Grue Problem," in Stalker, ed., pp. 225–40.

within scientific data analysis, my discussion will be very informal. In fact, my point of departure is a proposed solution, nearly thirty years old, that is a paradigm of an "analytic philosophy" approach to the problem, one that uses no tools from science or technical philosophy of science. This is Frank Jackson's proposal, introduced in 1975 in this JOURNAL and revised in 1980 with Robert Pargetter.[4]

Jackson argues that to deal with Goodman's problem we must explicitly attend to the fact that certain objects, the ones we are using as the basis for our induction, have the property of *having been observed* (or *examined*, or *being within our sample*). I will use the symbol "*O*" for this property. Different treatments of induction, and different inductive arguments, will involve slightly different *O* properties. In this article, I will abstract away from those differences as much as I can (though see Jackson, *op. cit.*, for more detail). So we start with a representation of inductive arguments that is something like this:

(1) All *F*s that are *O* are *G*.

(2) All *F*s (whether or not they are *O*) are *G*.

Goodman's problem shows that not all inductions that have the above form are good ones. Here we use the following definition of "grue."

Grue $=_{df}$ (green & *O*) or (blue & not *O*)[5]

Substituting "emerald" for *F* and "grue" for *G*, we create a bad argument. But Jackson in 1975 suggested that if we add one more premise (which does not involve a projectibility constraint on predicates themselves), the resulting form will always make for a good inductive argument. We need to add the "counterfactual condition." Rephrasing his proposal slightly:

(1) All *F*s that are *O* are *G*.
(2) If those *F*s were not *O*, they would still have been *G*.

(3) All *F*s are *G*.

This condition is met for *G* = green and not for *G* = grue.

[4] See Jackson, "Grue," this JOURNAL, LXXII, 5 (March 13, 1975): 113–31 (reprinted in Stalker, ed.); and Jackson and Pargetter, "Confirmation and the Nomological," *Canadian Journal of Philosophy*, X (1980). 415–28.

[5] This definition of "grue" is based on Jackson's 1975 discussion. Jackson also discusses various other definitions of "grue" in the literature, and shows why several of them (especially those that require particular grue emeralds to change from green to blue) fail to generate real philosophical problems about induction. Jackson's most precise definition of "grue" makes allowance for the fact that particular objects can change color. Below, *T* denotes some particular date, like January 1, 2020, while *t* is a variable. Grue at $t =_{df}$ (green and *t* & observed by *T*) or (blue at *t* & not observed by *T*)

Jackson's introduction of counterfactuals would be rejected as a nonsolution by Goodman himself. For Goodman, the concepts of law, disposition, natural kind, and counterfactual dependence form an interdefined network and are all dubious when interpreted in a strongly realist way. In response to worries of this kind, Jackson emphasizes that what we need here is knowledge of counterfactuals about the particular objects in the sample, not about emeralds in general.[6] I urge the reader to accept the use of these counterfactuals for a moment, and let the rest of the picture unfold. We return to the issue later.

Jackson's 1975 proposal works well for the case of "grue." But consider a different class of problem cases, "emerose" cases.[7]

$\text{Emerose}_1 =_{df}$ (emerald & O) or (rose & not O)

Now consider this argument:

(1) All emerose_1s that are O are green.
(2) If those emerose_1s were not O, they would still have been green.

(3) All emerose_1s are green.

Again we seem to have true premises and a bad induction. For a family of reasons, including reasons of this general kind, Jackson refined his proposal in his 1980 article with Pargetter. Their modification was to treat the F term and the G term similarly. I again rephrase their proposal.

(1) All Fs that are O are G.
(2) If those Fs were not O, they would still have been F and G.

(3) All Fs are G.

This will block the emerose_1 case. But consider a new "emerose" predicate.

$\text{Emerose}_2 =_{df}$ (emerald) or (rose & not O)

Now we do not take *any* emeralds out of the class of emeroses. This restores the problem for Jackson and Pargetter. If we substitute "em-

[6] Laura Schroeter (in discussion) suggested that Jackson's specific counterfactual claims might be resisted. Suppose we accept Goodman's arguments that similarity judgments (like claims about patterns and regularities) are language-dependent. But counterfactual claims, on many views, are dependent on similarity judgments. Then a speaker of a language in which "grue" is linguistically normal might be entitled to insist that if our particular observed emeralds had not been observed they would still have been grue, so they would also have had to have been blue. This objection raises interesting issues but I will not pursue it further here.

[7] See Goodman, p. 74, footnote 10.

erose$_2$" for *F*, then the second premise in the induction is true. But the induction is bad.[8]

A reply to my emerose$_2$ argument was proposed by Alexis Burgess (while an undergraduate, in a midterm).[9] Here is a modified version of his suggestion for the proper form of an inductive argument:

(JPB) (Jackson-Pargetter-Burgess)

(1) All *F*s that are *O* are *G*.
(2) There exists no property *C* (other than *G*, and weakenings and strengthenings of *G*) such that:
  (2.1) All *F*s that are *O* are *C*,
  (2.2) Some *F*s are not *C*, and
  (2.3) The *F*s that are *O* are *G* because they are *C*.

(3) All *F*s are *G*.

This deals with the emerose$_2$ case, and with the other cases as well. In the case of emerose$_2$, the relevant *C* is being an *emerald*. In our sample of emerose$_2$s there are only emeralds, and we have reason to believe that the greenness of these objects was due to their being emeralds. The emeraldness of the emeroses in our sample is interfering with the induction. The same applies in the case of emerose$_1$. In the case of the original grue problem, *C* is being *observed*. Here Burgess's proposal reduces to Jackson's original one.[10]

There are problems of detail with this proposal; it is not clear how best to formulate the second premise. In my formulation I used a "because" in (2.3) rather than an explicit counterfactual. This "because" is meant to include a variety of dependence relations, including both causal and logical ones. I emphasize, following Jackson, that this "because" claim is about the specific objects in our sample, and it should not be interpreted as involving too strong a brand of responsibility. The bracketed restriction on values of *C* can be ignored for now. I use the term *JPB proposal* (Jackson-Pargetter-Burgess proposal)

[8] This is not a case where one sample generates two conflicting inductions, and Jackson did require this in his 1975 article for there to be a philosophical problem, as opposed to a case of the ordinary fallibility of induction. But this is a case where green emeralds can be used to project greenness in all unobserved things, and similarly irrelevant observations can generate conflicting color projections about the same unobserved objects. So induction has collapsed as a way of discriminating good from bad predictions.

[9] Burgess has not published a discussion of this topic of his own (see his manuscript, "It's Not Easy Being Grue"). His view of the status of the proposal, and its best formulation, differ from the view defended here.

[10] Emerose$_1$ could also be dealt with by requiring via a different version of premise (2.3) that the sampled objects not be *F* (as well as not be *G*) because they were *C*. See the discussion of emerose$_1$ in the next section.

for a family of slightly different formulations that have the same basic structure. Different formulations and technical complications are discussed below in the appendix.

The most important point here is not the details of how to formulate the JPB proposal. What is important is that with Burgess's refinement of Jackson and Pargetter in place, we are in a position to see what is really going on in the grue/emerose family of problems. What Burgess has done is (inadvertently) assimilate the grue problem to a familiar set of issues in scientific methodology.

III. CONFOUNDING VARIABLES, HAWTHORNE EFFECTS, AND BIASED SAMPLES

In statistics and data analysis, one thing that people look for is associations between properties. But not all associations are regarded as genuine, in a sense that allows extrapolation, generalization, or causal interpretation. These associations are sometimes called *spurious*. How can one distinguish spurious associations from genuine ones? Sometimes one cannot, or not until things go wrong. But there are some well-known sources of error that can be guarded against. These problems are treated as defeaters of inferences that are ordinarily sanctioned.

The philosophical problems we have discussed here correspond to known sources of trouble in data analysis; they correspond to two sources of spurious associations. For a moment, it will be helpful to shelve "emerose$_1$" and focus on "grue" and "emerose$_2$."

The problem raised by "grue" is an unusual version of, or a close relative of, the problem of *confounding*. Here I have to stretch existing scientific terminology somewhat. The term "confounding variable" is usually only employed in discussions of causal inference. Judea Pearl formulates the most basic idea here as follows: if we are trying to work out whether there is a causal relation between X and Y, a confounding variable is another variable Z that affects both X and Y.[11] I prefer the following formulation: if we are trying to work out whether a variable X causally affects another variable, Y, a confounding variable is another variable Z that (i) affects Y, (ii) is nonrandomly associated with X, but (iii) is not itself causally dependent on X.

Suppose we want to know whether smoking causes heart disease. But when we do a survey of the population we find that the overall rate of heart disease is *lower* in the smokers. What does that show? We do not know; it might be due to a confounding variable. The

[11] See Pearl, *Causality: Models, Reasoning, and Inference* (New York: Cambridge, 2000), chapter 6

smokers might tend to be athletic people who exercise a lot. Then we might find that the overall rate of heart disease is lower in smokers than in nonsmokers, but as a consequence of a confounding variable, exercise. We can correct for the problem by dividing the population into exercisers and nonexercisers, and assessing the association between smoking and heart disease within each. It may turn out that within each group, smoking and heart disease are positively correlated, even though in the whole population the association disappears. (This is a case of "Simpson's paradox.")

So when doing causal inference we must guard against confounding variables. Confounding variables give rise to spurious associations (or spurious absences of association).[12]

The grue problem appears at first to be a very different kind of problem from the smoking/exercise problem. Here we are not doing causal inference, but pure projection. We are not trying to find out whether emeraldness causes greenness or vice versa, but only trying to find out whether all emeralds are green. We would not mind if there was a common cause of emeraldness and greenness, or if one caused the other, so long as all emeralds are green. But if we look at these "pure" projections closely, we find some kinship with the issues arising in causal inference. We should not make a projection from a sample if there seem to be the *wrong kind of dependence relations* between properties of the sampled objects.[13]

In the grue case, we find an association between emeraldness and grueness in our sample. But it is clear that this association has arisen for the wrong reasons. All the objects in the sample are observed things, as well as grue emeralds, and we know that observation *affects the application* of the predicate "grue." It is not that observation physically affects the objects in the sample, but it affects the ways that objects fall into classes expressed by predicates. This is a peculiar kind of dependence relation between features of the sample, which interferes with the projection (a kind of "Cambridge dependence"). It is epistemologically similar to the situation we would have if we had reason

[12] Pearl (*op. cit.*) argues that confounding is an irreducibly causal concept that cannot be expressed in statistical terms, but is also fundamental to data analysis in many sciences. The question of whether or not confounding can be defined in purely statistical terms is orthogonal to my concerns in this article.

[13] In traditional philosophical discussion, a distinction is drawn between the "observed" and the "unobserved" emeralds. In discussing the grue problem using the tools of scientific data analysis, the distinction that is more significant is that between the emeralds in a sample and the emeralds in the rest of the emerald population—the "sampled" and "unsampled" emeralds. To keep this article readable, I have moved between these distinctions without giving them much explicit attention.

to believe that the collection of the emeralds in our sample had included a process that painted or dyed those emeralds.

The problem with "grue" is similar to a confounding variable problem, as it involves the interfering role of an extra dependence relation that creates a spurious association. I do not claim that this is exactly the same as a classical problem of confounding, but the kinship is quite close. In causal inference, confounding involves a dependence relation that may create a spurious association in an entire population, misleading us about causal structure. In the case of projection (or estimation), the dependence between grueness and observation creates a spurious association in our sample, misleading us about the larger population of emeralds. Causal inference must guard against the possibility that it is not just the smoking that is affecting the health of smokers in our population, but unusual rates of exercise as well. Estimation and projection must guard against cases in which it is not just the emeralds' physical nature that is affecting how color predicates apply to them, but the process of observation as well.

A more specific comparison can also be made between the grue problem and what is known in the social sciences as the "Hawthorne effect." In a case of the Hawthorne effect, the subjects of a social scientific study behave differently because of their interaction with the observer studying them. They may work more efficiently, for example, as was conjectured in the case that generated the concept, a 1920s study of a factory in Hawthorne, Illinois.[14] In a Hawthorne effect case, O has a causal impact on other properties of interest in the study; in the grue case, O causes trouble via a noncausal dependence.

So my argument is that standard methodological principles in science tell us not to extrapolate grueness from a sample of grue emeralds, because we have reason to believe that the association in the sample is the product of a "bad" dependence relation that behaves similarly to a case of confounding, or a Hawthorne effect. Roughly, the semantics of "grue" turn observation itself into a confounding variable.

The emerose$_2$ problem is different. Here we do not find a bad *dependence* relation between properties of objects in the sample.[15] The emerose$_2$ predicate introduces a *selection bias* in our sample.

If we are to make a good inference using the basic techniques of

[14] The original study (which is now quite controversial) is discussed in E. G. Mayo, *The Human Problems of an Industrial Civilization* (New York: Macmillan, 1933) Mayo's discussion tends not to distinguish between the effects of special kinds of work supervision and the effects of being studied *per se*. I am grateful to Art Owen for bringing the Hawthorne effect to my attention.

[15] That is why the emerose$_2$ problem was not handled by Jackson and Pargetter's 1980 proposal.

statistical analysis of samples, every member of the global population we are interested in should have the same chance of being found in the sample. Often this is not possible in practice, but then we at least should try to ensure that the properties that made some objects likely to end up in the sample are not associated with the property we are making a projection about (property *G*). But the only objects that can be found in a sample of emerose$_2$s are emeralds. Roses cannot be found within the sample and still be emerose$_2$s, because of the peculiar semantics of the predicate "emerose$_2$."[16] So it is impossible to collect an unbiased sample of emerose$_2$s. And we have good reason to believe that the emerald/rose difference is relevant to inferences about color.

So standard methodological principles tell us not to extrapolate greenness from a sample of green emerose$_2$s to the total population of emerose$_2$s, because we have reason to believe that any such sample will be contaminated by selection bias.

In the grue and emerose$_2$ problems we find two different ways of generating a spurious association: selection bias, and something akin to confounding. These problems are, I emphasize, somewhat different. In the case of grue, we may have gathered a perfectly good sample of emeralds, but these objects have been affected, with respect to how crucial predicates apply to them, by their being observed. In the case of emerose$_2$, we just have a bad sample.[17]

The case of emerose$_1$ is slightly different again. We can exclude the emerose$_1$ projection as a case of selection bias, as with emerose$_2$. But in this case we also find an inappropriate dependence between properties of the sampled objects and the fact of being observed, as in the case of grue. If the emerose$_1$s in our sample had not been observed, they would not even have been emerose$_1$s (assuming they would still have been emeralds in that case). These objects would have been outside the population being projected *to*, as well as being outside the sample being projected from, because of the behavior of the term "emerose$_1$." This is an especially unusual case.

Burgess's generalization of Jackson is useful because it captures, in a rough but very informative way, a common feature of the various

[16] Here, and elsewhere, I assume that the sampled emeralds are the same as the observed ones, to avoid having to define different "grue" predicates to use in slightly different cases.

[17] See also J Moreland's "On Projecting Grue" for an earlier attempt to deal with the grue problem (though not emerose problems) using the concept of randomness, applied within a Carnapian framework—*Philosophy of Science*, XLIII, 3 (1976) 363–77. Note also that while I use the term "bias" in a specific way here, the term can also be used more generally to refer to any (nonchance) source of spurious association.

kinds of problem discussed above. In any induction, we need to try to ensure that there is not some property of the sampled objects that *makes those objects special* with respect to the question being addressed. This is what "*C*" in the JPB formula is aimed at picking out. If there is some such property, we should not think that what we find in the sampled objects can be extrapolated to the unsampled objects.[18] The grue and emerose predicates devised by philosophers do not raise a new kind of problem for nondeductive inference—instead they raise existing kinds of problem in an unusual way. We must guard against bias and confounding when making inferences from samples—that much we already knew. But most problems of these kinds come from empirical features of the situation, from extra causes we have not controlled for, or flaws in our sampling methods; we must make sure that our emeralds did not come to us via a process that physically alters their color. "Grue" and the "emerose" predicates are unusual because they are terms that *create* their own problems with confounding and bias, when used in inductions. That is why grueness in an emerald sample does not have the same significance that greenness has in such a sample. There is *no* way to control for the "effects" (in scare quotes) of observation on the properties of our sample in these cases. A standard green-emerald induction is a risky thing, but at least our model of the inference does not tell us in advance that it is unreliable.

Before concluding this section, I should revisit the issues concerning counterfactuals and "because" statements in the JPB proposal and my discussion of confounding. For followers of Goodman, any use of these tools in a purported solution of the new riddle produces a subtle circularity in the analysis. I will not address the most general issues of that kind here. The view being proposed does require that we accept and make use of the idea of dependence relations among the properties of individual objects. These can be described in a variety of ways, and can be philosophically analyzed in a variety of ways. No particular theory of counterfactuals is assumed here, for example (though see footnote 6 above). The kind of talk about dependence

[18] A referee suggested that once the family of problems have been linked in this way, they might be further analyzed by linking them to the "reference class" problem familiar from earlier discussions of induction. This possibility can be seen most readily in the case of emerose$_2$, but might be extended to the others. For example, *observed emeralds* are a narrower reference class than *emeralds*, and the peculiarities of the "grue" predicate might prevent us from using the frequency of grueness in the narrower class as a guide to the frequency in the broader class. The Hawthorne effect might be treated the same way. This suggestion raises interesting issues, but would involve recasting the discussion in a different framework and I will not follow it up further here.

relations that is needed for the solution to work is, as I hope this section has shown, a fairly down-to-earth kind that is familiar within science. It is a fairly extreme manifestation of Humeanism to deny that *any* dependence concepts of the kind required are legitimate.[19]

My proposed solution to Goodman's problem is not to be identified with the formula labelled "JPB" above. As we will see in the Appendix, it is hard to get a version of the JPB formula to deal with all the cases in the right way. And formulas like JPB, even when they seem to capture the cases correctly, do no more than capture a set of intuitions about justified inference. They do not say why this kind of transition from observed to unobserved is justified or reliable. Indeed, I do not think that JPB, or anything similar to it, represents a rule or argument form that is justified in all cases. What JPB does is abbreviate some key features of a more detailed treatment of reliable inference from samples, a treatment developed in statistics and various parts of science itself. That approach does more than capture intuitions; it uses a model of the sampling and inference process to show that some inference procedures are reliable (if the assumptions in the model hold). I do not deny that there are foundational problems with these aspects of scientific methodology. But I argue that this body of theory and practice does contain tools that resolve the grue and emerose problems, in the versions discussed above. The JPB formula, in turn, abbreviates the features of that theory that give us the crucial epistemic distinction between "grue" and "green" as they function in projections.

IV. THE LIMITS OF THE SOLUTION AND THE DISUNITY OF THE GRUE PROBLEM

I have argued that the restrictions needed to exclude the grue and emerose inductions already exist, implicitly, within the toolkit of scientific data analysis. However, I have developed this argument by forcing the grue induction into the form of one particular kind of inference found in science: inference from a sample to a larger population from which the sample is drawn, using statistical tools. This is, in many ways, the appropriate scientific case to focus on first. But as we will see, we need to think differently about grue when dealing with some other kinds of inference. Goodman's compact original discussions raise problems for more than one kind of inference, and it is a mistake to impose a single solution on what is really a collection of problems.

[19] One theme of Pearl's book is a defense of the scientific respectability of an anti-Humean, causal realist position, which is friendly even to counterfactuals. His treatment of confounding as a causal concept is part of that defense. My discussion here is not committed to the particular version of anti-Humeanism that Pearl defends

I will begin this part of the discussion by looking at some issues connecting the grue problem with laws and causation. The solution discussed in earlier sections of this article asserts no connection between the confirmation of generalizations and "lawlikeness." If all emeralds are "accidentally" green, this makes no difference to the story told above. It also does not matter whether or not the emeralds form a "natural kind." What matters is that the emeralds form a collection of objects that can be randomly sampled.

It does make a difference if the mechanisms responsible for color in the sampled emeralds are *different* from the mechanisms responsible for color in the unsampled emeralds. If we believe there is diversity in mechanisms that correlates with membership in the sample, we have a problem of the "confounding" kind discussed earlier. Whatever the facts are concerning the link between F and G in the sample, we need to assume that the same factors are operating in the unsampled part of the population, for a reliable inference to be made about the total population using the simple statistical model. What is making our inference procedure work in this case is the power of random sampling, as described by probability theory—the tendency for random samples to resemble the larger populations from which they are drawn. That is what is giving us the ability to move from the observed to the unobserved.

Although the distinction between natural kinds and mere collections, and between lawlike and "accidental" generalizations, does not matter in this context, the power of random sampling to connect the observed to the unobserved does depend on stringent conditions. We do not have to push very hard in order to push up against the limits of the set of inferences that can be treated that way, and hence to the limits of the proposal I made about Goodman's problem in the previous sections.

Think once again about emeroses. In the case of emerose$_2$s, only emeralds can be in our sample and still be emerose$_2$s. So there cannot be a good random sample of emerose$_2$s. But this points us towards a more general problem: *future* individuals cannot ever be in a present-day sample. Emeralds (or smokers, or ravens) that do not yet exist cannot find their way into our present-day samples. This shows us how quickly we reach the limits of the class of inferences that can be modeled and justified using the idea of random sampling alone. We can make an inference from a sample to a larger population without worrying about what sort of class, collection, or kind the larger population constitutes. But when the population from which we draw our sample is really a mere *sub*population of a larger population that we wish to project to, then we need substantive assumptions about

similarity between the subpopulation and the larger population. Otherwise, we have no reason to believe our sample is a good basis for projection to the larger population.

The epistemic gap between random sample and total population is fundamentally different from this gap between a subpopulation and a larger population "attached" to it. If we want to make inferences about a larger population that cannot be properly sampled, we must ask: What *kind* of collection is this? Are these objects all the products of a common type of origin? Do they have a common internal structure? What sort of causal or nomic relationship is there likely to be between properties we are projecting from and properties we are projecting to? Here we reach the kinds of issues usually discussed by philosophers in relation to "natural kinds."

So now we must ask a fresh set of questions about grueness. Once again we imagine having before us a pile of emeralds, but now we cannot see this pile of emeralds as drawn from a total population by random sampling. Instead, our pile of emeralds is like a little subpopulation attached to the rest of the emeralds (or a sample of such a subpopulation). Once again, we ask why we should not project grueness to the rest of the emeralds. In this context, Goodman's argument is raising a different set of issues from those discussed earlier. Now it is a very important fact about grue that it is an odd-looking property in relation to our background knowledge for the case at hand.

Here we do run into inconvenient features of the standard example. Minerologists seem to treat the greenness of emeralds as a defining characteristic of the kind: emeralds are beryl crystals made green by chromium impurities. So let us stipulate that the kind being investigated is beryl crystals containing chromium impurities. The focus of our investigation will then be the way a beryl crystal with chromium responds to light. As the grammar of my previous sentence suggests, in this investigation it is only practical considerations that make a sample of a hundred emeralds superior to a sample of one. The mere addition of new emeralds to our data set does not, as it does if we are making an estimate from a random sample, make much difference to the situation.[20]

[20] The debate between John Dewey and Hans Reichenbach in Dewey's "Schilpp volume" includes an interesting exchange on this issue. Reichenbach modeled all nondeductive inference on statistical estimation, and hence saw the role of sample size in generating convergence on a true value as crucial. Dewey had a different model of nondeductive inference, in which all hangs on the ability to find an individual that is representative of its kind; if we can do this, then one individual is enough. I am suggesting that both Dewey and Reichenbach were on the right track with respect to understanding some inferences in science, but both were too inclined to general-

I will not try to give a detailed account of exactly which factors bear on this form of the grue problem. Clearly though, the large literature on natural kinds is relevant here.[21] Some of the other tools employed in more standard discussions of grue—parsimony, prior probability—may also be important. I would favor a multi-factored treatment of this aspect of the problem. The special features of "grue" that Jackson focused on still play a role. When we make inferences about emeralds from those we have seen, it is relevant to ask whether there are any likely effects on our emeralds of the processes by which they made their way into our possession. But a great deal of attention must be paid to something that has no role in the JPB approach—what sort of kind *emerald* is, and the network of properties (especially intrinsic, structural properties) characteristic of this kind.

So the proper treatment of grue in a purely statistical inference is different from its proper treatment in an inquiry into the causal and nomic properties of a putative natural kind. This disunity is to be expected; we have here two rather different kinds of inference found in science. The two kinds of inquiry are often connected, and they may sometimes combine so closely that they are hard to disentangle in a particular case. But they do raise different epistemological issues.

We see an indication of all this in the JPB formula itself. We ask: Is there a particular known factor *C* that explains the *G*-ness of the sampled *F*s? If no, we proceed with the projection, though of course we might be wrong. If the answer is yes, there are two options. We might believe that *C* is not found in the unsampled *F*s. In that case, *C* is a kind of confounding factor; we cannot proceed with the inference. But if factor *C* *is* found in all the unsampled *F*s, then it seems that we have something close to a *deductive* argument for the conclusion that all *F*s are *G*. We do not need to rely on our random sample any more. Further questions do arise; maybe we cannot rely on *C* to have the same effects in the unsampled cases...? But to say this is to embark on an entirely different kind of investigation from the statistical inference we started with.

These considerations have consequences for the way that philosophy has formulated and addressed the problems surrounding "induction." In this section, I have emphasized the differences between two

---

ize. See P.A. Schilpp and L.E. Hahn, eds., *The Philosophy of John Dewey* (Library of Living Philosophers, La Salle, IL: Open Court, 1939).

[21] For the bearing of natural kinds on induction, see especially W.V. Quine, "Natural Kinds" (in *Ontological Relativity and Other Essays* (New York: Columbia, 1969), reprinted in Stalker, ed.); and also Hilary Kornblith, *Inductive Inference and Its Natural Ground* (Cambridge: MIT, 1993).

distinct kinds of scientific investigation and inference. One is statistical inference from samples. The other is inference about the structures and mechanisms responsible for the clustering of properties in a kind. Clearly these are not the only two kinds of investigation in science, but they are part of the story.

Statistical inference from samples is a natural tool for investigating the relation between two variables when the causal structure relating them is complex, mixed, and imperfectly known. Once we start to develop a reasonably unified story about what sort of causes operate in *F*s, and how these relate to the property *G*, it will be natural to move to a different kind of investigation. Then the multiplication of instances does not much matter. Inferences are based not on the power of random sampling, but on tracing specific dependences and causal paths.[22] We may introduce a new set of categories to organize the domain under discussion, too, when we switch methods; we may no longer treat "the *F*s" as the right class for analysis.

The philosophical concept of "induction," especially in the years after Goodman's discussion, seems often to be a mixture of these two separate kinds of scientific inference. Science contains one kind of inference where sample size *is* important and "naturalness" of kinds is *not*. In these inferences, it is the power of random sampling that gives us our link between observed and unobserved. Science also contains a second kind of inference in which sample size is *not* very important, but the status of the kinds under discussion *is*. Roughly speaking, it is the causal reliability of structures and mechanisms that gives us our link between the observed and unobserved. (Once again, I do not claim that these exhaust the inference patterns found in science.) The philosophical concept of induction in much recent discussion, however, *includes* the idea that the number of observed cases matters, does *not* include an explicit role for randomness of sampling, and *includes* a role for the "naturalness" of kinds. The result is a hybrid form of inference in which it is supposed to matter both how many *F*s you have seen *and* whether *F*s are a natural kind. But the reliable operation of inner mechanisms common to a natural kind and the randomness of sampling are *two distinct routes to projection*; two distinct ways in which the observed cases can connect us to the unobserved.

[22] In biology, the much-debated statistical concept of *heritability* is an example (see R.C. Lewontin, "The Analysis of Variance and the Analysis of Cause," *American Journal of Human Genetics*, XXVI (1974): 400–11). Heritability is slowly becoming less important to genetics, as particular causal pathways linking genes and phenotypic traits are uncovered in more detail.

I will summarize my main points. Jackson argued in 1975 that all predicates are projectible in principle, so long as inductions respect his "counterfactual condition." He argued that giving a theory of the projectibility of predicates is the wrong way to address Goodman's problem. The first part of my article is in partial agreement with Jackson's claims. In the context of statistical inference from the properties of a sample to the properties of a larger population, any predicate can be projected from a random sample, provided that the inference is not affected by biased sampling, confounding (in my broad sense of the term), or similar problems. However, grue-like predicates tend to create, via their peculiar semantics, problems of just these kinds.

Goodman's problem also arises for a quite different kind of inference, which should not be conflated with the genuinely statistical cases. These are inferences about the characteristic features of a natural kind, generated by the causal and nomological mechanisms operating in that kind. The treatment of Goodman's problem in this second class of cases does still involve considerations of the sort Jackson described, but includes much more besides. It is a mistake to conflate these two distinct kinds of inference by positing a single category of "inductive" arguments, in which both sample size and "naturalness" matter.

PETER GODFREY-SMITH
Research School of Social Sciences/Australian National University
Harvard University

APPENDIX: FORMULATIONS OF THE JPB PROPOSAL

I will briefly discuss different formulations of the JPB proposal. The issue is not critically important, because I regard the JPB proposal not as itself the solution to (part of) Goodman's problem, but as an abbreviation, in the language of philosophical discussions of induction, of a treatment found in statistics and science. So I am uncertain about the significance of technical problems with JPB.

Here is my initial formulation of the JPB proposal.

(JPB)
(1) All *F*s that are *O* are *G*.
(2) There exists no property *C* (other than *G*, and weakenings and strengthenings of *G*) such that:
  (2.1.) All *F*s that are *O* are *C*,
  (2.2.) Some *F*s are not *C*, and
  (2.3.) The *F*s that are *O* are *G* because they are *C*.

(3) All *F*s are *G*.

The main question is how best to formulate premise (2), especially clause (2.3). I used a "because" formulation. Other possibilities include using an "only because" formulation, using an explicit counterfactual, and using conditional probabilities. One source of trouble is that premise (2) can be made so strong that it results in the argument form being either deductively valid, or close to it, and hence a bad representation of good "inductive" arguments.

That problem is what motivates, for example, the restriction against weakenings and strengthenings of *G*, as permissible values of *C*. The following objection was given (in slightly different form) by Aldo Antonelli (personal communication). If we allow *C* to be a weakening of *G*, then, at least under many versions and construals of premise (2.3), the JPB argument form is deductively valid. This problem is clearest for versions of (2.3) that use an explicit counterfactual. Suppose we used the following:

(2.3*) If the *F*s that are *O* had not been *C*, they would not have been *G*.

Then if we are allowed to consider *C*s that are weakenings of *G*, one candidate for *C* is $O \vee G$. Clauses (2.1) and (2.3*) are true for that choice of *C*. But then for premise (2) as a whole to be true, clause (2.2) must be false for that *C*. Clause (2.2) says that some *F*s are not *C*, so in the case considered here, this will only be false if all *F*s are $O \vee G$. But given that we also know that the *F*s which are *O* are *G*, all the *F*s (observed or unobserved) must be *G*. The JPB premises deductively imply the truth of the conclusion.

So if premise (2) is formulated in a way that uses or implies a counterfactual of that kind, we need a restriction on values of *C*. Similar problems arise if we allow candidate *C*s that are strengthenings, rather than weakenings, of *G*. (Consider $C = F \,\&\, G$.) In the main text I used a "because" formulation that is easy to grasp in many cases, but which is also less precise and may have an uncertain relation to the counterfactuals. In any case, it should be intuitively clear why the need for a restriction arises. The point of the JPB formula is to block inductions in cases where some extra dependence relation is affecting the *G*-ness of the objects in our sample. But various near-relatives of *G* (not to mention *G* itself) will have close logical and counterfactual connections with *G*. These near-relatives can act as "pseudo-explainers" of the *G*-ness of the sampled objects. We do not want these to be considered as possible defeaters of an induction, so they are ruled out as relevant values of *C* in the JPB formula.

Another family of problem cases is due to John MacFarlane and Branden Fitelson (personal communication). Let $C = M_1 \vee M_2 \ldots \vee M_\infty$ where this is a disjunction of all the very specific mechanisms

responsible for the *G*-ness of the *F*s in the sample. The threat is that reasonable-looking inductions will be disallowed because this sort of disjunction will always (or very often) provide a *C* that passes the test in the second premise of the JPB formula. I suggest, however, that if a *C* of that kind is put forward in an abstract, stipulative way, without adding more specific information about the case, there is no reason to believe that (2.3) is true. If the sampled *F*s had not been subject to the operation of $M_1 \vee M_2 \ldots \vee M_n$ who knows what they have been like? But this discussion of candidate *C* properties that involve causal mechanisms also takes us back to issues raised in the final section of the main text. Once we start to develop detailed individual explanations for why *F*s in the sample are *G*, then all hangs on whether *F*s outside the sample are subject to the same explanatory factors as those inside. If we have confidence that the explanation for the *F*s inside the sample being *G* holds elsewhere, then we do not need random sampling to support the conclusion that all *F*s are *G*; we can make a direct argument from the operation of these explanatory factors. But if the *F*s outside the sample are subject to different influences from those inside, in this respect, then we should not project from our sample at all.

A distinct problem is how strong the dependence relation expressed in (2.3) should be. It might be advisable to use a formulation that is explicitly weaker than the "because" I used. A "might" counterfactual is one possibility, but it may instead be better to move to an explicitly probabilistic formulation.

THE METAPHYSICS OF REALIZATION, MULTIPLE REALIZABILITY, AND THE SPECIAL SCIENCES*

What I shall term the "received" view of *multiple realizability* (henceforth MR), and also the special sciences and their laws, was originally laid out in the seventies in a famous paper by Jerry Fodor[1] and has since been elaborated by many others.[2] As befits its hegemonic status, a growing number of writers have recently criticized the received view.[3] Rather oddly, much of this work has proceeded without precise criteria for either the realization relation or multiple realizability itself. A significant exception, however, is the work of Lawrence A. Shapiro[4] who offers a range of important examples from which he derives a prima facie plausible criterion for MR and hence argues to some striking conclusions. For example, Shapiro argues against the view of MR outlined in a prominent paper

* For comments, I would like to thank my commentators Antony Dardis, at the American Philosophical Association, and Sara McGrath, at the Society for Philosophy and Psychology, and the audiences at the following venues where I read versions of this paper in 2003: the Pacific APA Conference in San Francisco; the SPP at Caltech; and the Joint Sessions in Belfast. I am especially grateful to Larry Shapiro for helpful discussion.

[1] "Special Sciences," *Synthese*, XXVIII (1974): 97–115.

[2] As will become clear below, I take the "received" view to be an integrated position with commitments about the nature of realization, MR, and the special sciences and their laws. The specific character of these claims will be clarified in sections I–III. Variants of the received view have been elaborated by a number of philosophers. Recent examples in the philosophy of mind are Terry Horgan's "Non-reductive Materialism and the Explanatory Autonomy of Psychology," in Steven J. Wagner and Richard Warner, eds., *Naturalism: A Critical Appraisal* (Notre Dame: University Press, 1993), pp. 295–320; and Robert Van Gulick's "Who's in Charge Here? And Who's Doing All the Work?" in John Heil and Alfred Mele, eds., *Mental Causation* (New York: Oxford, 1993), pp. 233–58. In the philosophy of biology, see, for instance, Harold Kincaid's "Molecular Biology and the Unity of Science," *Philosophy of Science*, LVII (1990): 575–93; and Philip Kitcher's "1953 and All That: A Tale of Two Sciences," *Philosophical Review*, XCIII (1984): 335–73. And, for another, rather different, defense of a variant of the received view against Shapiro's critique (note 4 below), see Alex Rosenberg's "On Multiple Realization and the Special Sciences," this JOURNAL, XCVIII, 7 (July 2001): 365–73. (The reasons why I do not follow the usual approach and include the work of Hilary Putnam, perhaps the earliest proponent of MR, as an exemplar of the received view of MR will become clear in section I).

[3] See, for example, Robert Batterman's "Multiple Realizability and Universality," *British Journal for the Philosophy of Science*, LI (2000): 115–45; William Bechtel and Jennifer Mundale's "Revisiting Multiple Realization," *Philosophy of Science*, LXVI (1999): 175–205; Lenny Clapp's "Disjunctive Properties: Multiple Realizations," this JOURNAL, XCVIII, 3 (March 2001): 111–36.

[4] "Multiple Realizations," this JOURNAL, XCVII, 12 (December 2000): 635–54.

0022-362X/03/0011/591–603 © 2003 The Journal of Philosophy, Inc.

by Ned Block and Fodor,[5] and the defenders of the received view following them. Contrary to Block and Fodor, Shapiro contends that cases such as the mammalian eye and octopus eye are *not* examples of multiple realization. Furthermore, building on his conclusions about MR, Shapiro defends a revisionary account of the special sciences and their laws which again challenges the received view.

Despite the importance of his examples and criterion, it is noteworthy that, like other recent critical discussions, Shapiro does not offer a precise account of the realization relation. And it appears plausible that providing such an account of *realization* would be a necessary step in any precise understanding of *multiple* realization. My main goal here will be to argue that this oversight is indeed a damaging one. In section I, I shall distinguish two opposing views about the metaphysics of realization. I will then show, in section II, that Shapiro implicitly assumes a metaphysical account of realization that is in conflict with the view held by his opponents over multiple realization. As a result, I shall argue that, although we should endorse Shapiro's criterion for MR, his criticisms of the received view based upon this criterion beg the question against his opponents.

Before concluding, in section III, I will also suggest that similar concerns vitiate Shapiro's wider claims about the nature of the special sciences, as well as those of another prominent critique recently offered by Lenny Clapp (*op. cit.*). For we shall see that differences over the metaphysical nature of realization are inextricably bound up with broader disputes over the extent of MR and the nature of the special sciences themselves. It is worth briefly marking my goals at this point, for I am obviously not attempting to resolve the many outstanding issues surrounding multiple realizability. Rather, I will simply show that philosophers of psychology must pay careful attention to the metaphysics of realization in their research and that a failure to do so undermines recent critiques of the received view.

I. TWO VIEWS OF THE METAPHYSICS OF REALIZATION

In his article, Shapiro focuses on the realization of "kinds" and does not explicitly adopt the causal theory of properties. However, I contend that conclusions derived using the causal theory of properties can be translated to such "kinds." Given this point, I shall therefore follow the recent metaphysical debate over realization and assume a

[5] "What Psychological States Are Not," *Philosophical Review*, LXVIII (1972): 159–81. I take the latter to be a seminal expression of the received view's commitment to an extensive view of MR.

causal theory under which a property is individuated by the powers it contributes to individuals.[6]

The first metaphysical view I want to outline is what I shall term the *flat* account. This view has been the main focus of recent work to clarify the metaphysics of realization and has been defended by both Jaegwon Kim and Sydney Shoemaker, among others. The flat view comprises two interconnected claims. The first thesis concerns the individuals in which realizer/realized properties are instantiated and can be framed thus:

(F-I) A property instance *X* realizes a property instance *Y only if X* and *Y* are instantiated in the same individual.[7]

The second thesis of the flat view concerns the causal powers of realized/realizer properties and may be captured as follows:

(F-II) A property instance *X* realizes a property instance *Y only if* the causal powers individuative of the instance of *Y* match causal powers contributed by the instance of *X* (and where *X* may contribute other powers not individuative of *Y*).[8]

I shall take any account of realization that incorporates (F-I) and (F-II) to be an example of the flat view. And through these two conditions the flat account earns its name by taking realized and realizer properties to share both the individual in which they are instantiated and at least some of the causal powers contributed to this individual.

A central feature of the flat metaphysics that bears emphasis is the quite literal notion of "causal role playing" that it embodies. Under the flat view one property instance realizes another only if the realizing property contributes *all* the powers individuative of the realized property—the realizer property instance thus literally plays the very causal role that individuates the realized property. As a consequence, the flat account makes no room for realization between properties in different individuals, for it appears that it is the very nature of instantia-

[6] In my discussion I shall use the version of the causal theory from Sydney Shoemaker's "Causality and Properties," in Peter van Inwagen, ed., *Time and Cause* (Boston: Reidel, 1980), pp. 109–35. To cover as many participants in the debate as possible, I will use a weakened form of this account which merely claims that in the actual world all instances of a property contribute the same powers under the same conditions.

[7] See Kim's *Mind in a Physical World* (Cambridge: MIT, 1998), p. 82, and p. 297 of Shoemaker's "Self, Body, and Coincidence," *Proceedings of the Aristotelian Society*, Supplementary Volume LXXIII (1999): 287–306.

[8] Kim, p. 54, and p. 78 of Shoemaker's "Realization and Mental Causation," in Carl Gillett and Barry Loewer, eds., *Physicalism and Its Discontents* (New York: Cambridge, 2001), pp 74–98.

tion that a property instance only contributes powers to the individual in which it is instantiated. If a realizer property instance matches the powers individuative of the realized property, the realizer must thus be instantiated in the same individual as the realized property.

In contrast, the second metaphysical view—what I term the *dimensioned* account—allows that realized and realizer properties may be either instantiated in the same, or in different, individuals. This account consequently differs in a range of other respects from the flat view—not least in its notion of "causal role playing." Under the dimensioned view, properties/relations F_1–F_n of the constituents of an individual s can play the causal role of a property G of s without F_1–F_n contributing any of the individuating causal powers of G. For in cases of realization involving different individuals, the realizer properties/relations instantiated in constituents do *not* contribute powers to the constituents that are individuative of the realized property G instantiated in the constituted individual. Nonetheless, F_1–F_n play the causal role of G in a wider sense, for the constituted individual has the powers individuative of G *in virtue of* the distinct powers contributed by F_1–F_n to the constituents.

We may more precisely frame the view of realization built around this wider notion of causal role playing using the following definition:

> (Dimensioned realization) Property/relation instance(s) F_1–F_n realize an instance of a property G, in an individual s, *if and only if* s has powers that are individuative of an instance of G in virtue of the powers contributed by F_1–F_n to s or s's constituent(s), but not vice versa.[9]

Unsurprisingly this definition is complicated since it encompasses cases of realization in different individuals, as well as those within individuals, and hence covers examples involving distinct individuals and powers, as well as different properties.

To better illuminate these opposing accounts, I want to examine briefly a concrete case over which they disagree, since this will also show how such views may be important in discussions of MR. As our example, let us use a disagreement Shapiro notes between the earliest

[9] The definition is articulated and defended further in my article, "The Dimensions of Realization: A Critique of the Standard View," *Analysis*, LXII (2002): 316–23, and also "Moving beyond the Subset Model of Realization" (unpublished manuscript). I should more carefully mark my intentions in offering the definition. First, I intend it to be an account of the notion of realization implicit in scientific theorizing, and not any folk concept. Second, I take realization to be a basic metaphysical notion whose nature is intimately bound up with a family of notions. I thus offer my definition of realization as an illuminating, nonreductive account of its connections to other basic metaphysical notions.

defenders of MR (*op. cit.*, pp. 645–46). Shapiro argues that Hilary Putnam does *not* take the mammalian eye and octopus eye to be a case of multiple realization, whereas Block and Fodor do class it as such.[10] Can we explain these divergent judgments? It appears we can. Both of these eyes operate using a single lens producing an inverted image that is then converted into electrical impulses. As a result, in each case the individual instantiating the property of being an eye has what Putnam calls a "virtually identical structure (physically speaking)."[11] Under the flat metaphysics, the only properties relevant to realizing a property instantiated in some individual *s* are other properties also instantiated in *s*. Thus apparently applying the flat view, Putnam denies we have a case of multiple realization, for the individuals instantiating the property of being an eye are "virtually identical" with respect to those of their properties which result in their instantiating the property of being an eye.[12]

Of course, the mammalian and octopus eyes do differ markedly in the kinds of cells, proteins, and other entities that compose them, but the properties of such constituents cannot be realizers under the flat view and hence cannot ground a case of multiple realization for

[10] See Putnam's "The Nature of Mental States," and "Philosophy and Our Mental Life," in his *Mind, Language, and Reality: Philosophical Papers*, Volume 2 (New York: Cambridge, 1975), pp. 429–40, and pp. 291–303.

[11] "The Nature of Mental States," p. 436.

[12] At this point, we should mark an idea recently broached by Sara McGrath, and others, about the potential role of structural properties as defined in David Armstrong's *Universals and Scientific Realism* (New York: Cambridge, 1978). McGrath's suggestion is that the two individuals instantiating the property of being an eye may be an example of multiple realization even under the flat account. The individuals instantiate distinct structural properties, given their different constituents, and the idea is that these structural properties provide different realizers of the property of being an eye. This suggestion requires more extended discussion than we can give it here in order to be properly evaluated, but in response it is worth briefly making two points. First, we should acknowledge that the questions raised by the suggestion are difficult ones given various unresolved issues about structural properties. For instance, there are open questions over particular matters such as the powers that the structural properties contribute to individuals. As well as more general problems, such as whether we should posit structural properties at all in addition to acknowledging the existence of certain constituents and their properties/relations. However, we cannot resolve these questions here, so having noted them, I shall put them to one side. The second point to emphasize is that proponents of the flat view such as Putnam or Shapiro will likely reject the suggestion. For they will argue that the powers contributed by the structural property of being made of certain proteins, that is, having a certain composition, comprise powers such as melting at a certain temperature or chemically reacting with specific substances. But these powers are apparently not powers in virtue of which the individual instantiates the property of being an eye, so Putnam or Shapiro will resist the conclusion that structural properties underwrite MR on the flat view. However, it again remains to be determined whether such a response to the suggestion is successful or not.

Putnam. However, anyone using the dimensioned metaphysics might potentially argue that this is a case of multiple realization, for under this metaphysics the properties/relations of the constituents of an individual are potential realizers of this individual's properties. On such a view, what is important is that the properties/relations of both types of constituent contribute powers that *result in* the powers individuative of having an eye in the pertinent constituted individual. Thus commitment to the dimensioned view is plausibly just what underlies Block and Fodor's judgment that this is a case of multiple realization.

Using recent work on the metaphysics of realization, we thus see that this early dispute over MR can arguably be traced back to a deeper implicit difference over the nature of the realization relation. In itself, this is a telling point about the importance of these metaphysical issues, for Putnam, Block, and Fodor are often lumped together as holding the same basic position. Bearing this cautionary point in mind, let us turn to Shapiro's arguments about multiple realizability.

II. UNDERSTANDING THE RELEVANCE CRITERION

The key example that Shapiro offers concerns two corkscrews that differ only in color and he plausibly argues that such a case does not constitute multiple realization of the property of removing corks. Shapiro says:

> To say that a kind is multiply realizable is to say that there are *different* ways to bring about the function that defines the kind. But, if two particulars differ only in properties that do not in any way affect the achievement of the defining capacity of a kind, then there is no reason to say that they are tokens of different realizations of the kind (*op. cit.*, p. 644).

Using the causal theory of properties we can capture the underlying idea here in what I shall call the *relevance criterion*, or the *criterion*, as follows:

> (Relevance criterion) Instances T_1, instantiated in s_1, and T_2, instantiated in s_2, are distinct realizations of a property T *only if* there is a difference between the properties/relations whose contribution of powers are those in virtue of which s_1 and s_2 have the powers individuative of T.

Here we can see that the criterion entails that we have a case of multiple realization only if there are differences in the properties/relations, which Shapiro calls the "causally relevant properties," whose contributions of powers result in the causal powers individuative of the realized property. The criterion thus frames the reason why the colored corkscrews are not an example of multiple realization, for

the property of having some color does not contribute powers that result in the individuative powers of the property of removing corks.

The relevance criterion plausibly advances our understanding of MR and with it in hand Shapiro turns to a range of putative cases of multiple realization. For instance, Shapiro considers two corkscrews that are otherwise identical except one is made of steel and the other of aluminum. Applying the criterion, Shapiro then argues as follows:

> Steel and aluminum are *not* different realizations of a waiter's corkscrew because, relative to the properties that make them suitable for removing corks, they are identical. The fact that one corkscrew is steel and the other aluminum is no more reason to characterize them as different realizations than the fact that one might be yellow and the other red (*op. cit.*, p. 644).

Initially, this application of the criterion might be puzzling to some, but it makes sense if Shapiro is assuming the flat view of realization in interpreting the criterion. For under the flat account, the only properties that can be realizers are instantiated in the same individual as the realized property. Shapiro thus argues that the properties/relations of the constituent steel molecules and aluminum atoms cannot be the basis for a multiple realization, since they are not even relevant to the realization of the corkscrews' property of removing corks. These properties are not instantiated in the corkscrew, but in its constituents, and using the flat interpretation of the criterion yields the conclusion that the steel and aluminum corkscrews are not different realizations.

Further evidence that Shapiro implicitly assumes the flat metaphysics in applying the criterion is provided by the manner in which he adjudicates the dispute between Putnam and Block and Fodor noted earlier. Shapiro argues as follows:

> ...Putnam might shake off as irrelevant the information that the lens of the octopus's eye is made of one kind of protein and the lens of the mammalian eye of another.... The [common] processes that result in the formation and analysis of an image screen off these facts about microcomposition.... For Block and Fodor, eyes would count as distinct [realizations] whenever they differ in molecular composition. But if the line of argument I have pursued here is correct, Block and Fodor would for similar reasons be committed to distinguishing between kinds of corkscrews simply on the basis of color (*op. cit.*, p. 646).

Shapiro ultimately sides with Putnam (*op. cit.*, p. 647), arguing that the criterion implies that differences in the constituents of the eyes have no bearing on whether the eyes are differently realized. As I earlier argued was the case with Putnam, Shapiro is apparently assum-

ing the flat metaphysics under which the only properties relevant to realization are instantiated in the same individual as the realized property. Furthermore, Shapiro then concludes that, given what he takes to be Block and Fodor's mistaken judgments in such cases, they can only be working under the assumption that *any* difference in properties suffices for MR. Shapiro thus concludes Block and Fodor must implausibly class the difference in color of the corkscrews to comprise a case of multiple realization. Consequently, Shapiro implies that we should accept a more modest view of the extent of multiple realizability than has been the received position in philosophical debate.

Having laid out Shapiro's key arguments about MR, let us now turn to evaluating his claims. As we saw in section I, Block and Fodor arguably implicitly assume the dimensioned, rather than the flat, metaphysics. Shapiro does not explicitly defend a particular account of realization, nor does he distinguish the flat and dimensioned views. Instead, like most others in the debate, Shapiro unknowingly assumes a disputed view of realization, in his case the flat view. An important question is consequently whether we still reach Shapiro's negative conclusions using the criterion when it is interpreted in light of the alternative model of realization plausibly assumed by Block and Fodor and other proponents of the received view. I will show that we do not.

To illuminate this point, let us examine some cases from the perspective of a defender of the dimensioned metaphysics, starting with the steel and aluminum corkscrews. The dimensioned view allows properties/relations of constituents to be realizers and Block and Fodor will plausibly take this to be a case of multiple realization. For with one corkscrew the properties/relations of steel molecules result in the powers individuative of the property of removing corks, while in the other corkscrew the distinct properties/relations of aluminum atoms play this role. We should carefully mark that this judgment accords with the criterion—for only properties/relations that result in the powers of the realized property are taken to be relevant to MR. The difference from Shapiro's use of the criterion is simply that the dimensioned view has been assumed with its wider understanding of both "causal role playing" and a realizer property.

Do proponents of the dimensioned metaphysics such as Block and Fodor, as Shapiro implies, therefore take *all* differences of composition to be instances of multiple realization? To see that they do not, consider two aluminum corkscrews that are similar in all other respects except that one is made of aluminum containing a trace element. This element does not chemically bond with the aluminum, or change the metallic structure of aluminum atoms, but it does

absorb a certain wavelength of light giving this corkscrew a yellow tinge. The same structure of aluminum atoms is therefore responsible for rigidity in both corkscrews, but there is a trace element in one of them.

Shapiro implies Block and Fodor must accept that such a case is a multiple realization of the property of removing corks. But charitably interpreting Block and Fodor as holding the dimensioned model, they will *reject* this as a case of multiple realization even though there is a difference in microcomposition. Applying the dimensioned view to this case, we see that in both corkscrews the realizers will be the same properties/relations of aluminum atoms. For the properties/relations of the atoms of the trace element do not contribute powers that result in the individuative powers of the property of removing corks and hence cannot either be realizers of the latter property or, consequently, the basis for distinct realizations of it. Once more these conclusions comport with the truth of the criterion, albeit seen through the lens of the dimensioned metaphysics.

The underlying point of these interindividual examples of realization also shows that Shapiro is mistaken in his claim that Block and Fodor must take an intraindividual case such as the colored corkscrews to be an instance of multiple realization. Again, we should immediately note that the property of being red, or some other color, will not be a realizer of the property of removing corks under the dimensioned account, since it does not contribute powers resulting in the individuative powers of the property of removing corks. Thus color is not a realizer of the property in question under the dimensioned view and Block and Fodor may argue that the difference in color provides no basis for the existence of distinct realizations even in this intraindividual case.

These examples establish that defenders of the received account of MR will *accept* Shapiro's important insights about the relevance criterion and its underlying idea that the only properties/relations relevant to MR are those in virtue of which the pertinent individual has the individuative powers of the realized property. However, as we have seen, there are nonetheless differences in the accounts of MR endorsed by Shapiro and proponents of the received view such as Block and Fodor, for they diverge over which properties can be realizers in any case due to their opposing views of realization. Given his implicit acceptance of the flat view, Shapiro only allows as relevant in applying the criterion those properties/relations literally playing the causal role of the realized property—that is, contributing the individuative powers of the realized property—and hence instantiated in the same individual as the realized property. In contrast, defenders of the

dimensioned view take a much wider class of properties as potential realizers and hence apply the relevance criterion to produce a view of MR that is wider in scope. Proponents of the received view consequently take the steel and aluminum corkscrews, and the mammalian and octopus eyes, as instances of multiple realization, though still endorsing the truth of the criterion and denying the colored corkscrews are distinct realizations.

We can therefore see that Shapiro's main example, and the relevance criterion drawn from it, do not provide a criticism of proponents of the received view of MR, such as Block and Fodor, *unless* one implicitly assumes an account of realization that they dispute. Shapiro's failure to provide an independent account of the realization relation thus apparently leads him implicitly to assume a contentious view of the nature of realization without being aware of this. And he therefore fails to see how his opponents may be charitably interpreted as accepting his main points on their own terms.

Furthermore, it should now be obvious that the dispute over the metaphysics of realization is not separate from the debate over the extent of multiple realizability. We have found that both sides should endorse the underlying idea of the relevance criterion, but are led to different applications of it simply by their opposing metaphysical views of realization. The question of the correct account of realization is thus inextricably bound-up with the issue of the nature of MR. If one therefore assumes a certain view of realization in arguing against an opposing position about MR one runs a very real risk of begging the question. As we have seen, Shapiro's key arguments, based upon his loaded interpretation of the criterion, plausibly do beg the question against defenders of the received account of MR in just this manner.

III. REALIZATION AND THE NATURE OF SPECIAL SCIENCES

Building upon his conclusions about MR, as I noted earlier, Shapiro also defends a range of further claims about the nature of the special sciences and their laws. It is worth briefly examining the dilemma Shapiro uses to derive these wider conclusions, for we are now in a position to evaluate it more carefully. If we accept his arguments about MR, Shapiro asks "why should we think that [the multiple realizability thesis] is a coherent thesis at all?" and uses the following dilemma to support this point (*op. cit.*, p. 647). Either realizers do, or do not, differ in their "causally relevant properties" from realized properties. But, on the first horn, if realizers do differ in their "causally relevant properties," then one will have no single realized property and hence no multiple realization. On the other horn, if realizers do

not differ in their "causally relevant properties," then one does not have distinct realizers and again we have no multiple realization. Thus we have no multiple realization! In response to this dilemma, Shapiro then outlines what he takes to be a legitimate role for the special sciences that slips its horns, albeit a role that differs markedly from the received account.

We can quickly see, however, that commitment to the flat metaphysics is necessary for any such worrying dilemma to arise. In this case, the work is apparently done by the flat view's assumption that only properties of an individual *s* that literally play the causal of a realized property of *s* can be its realizers. As a result, it is all too easy to conclude, as Shapiro apparently does, that if the same causal role is played by the relevant realizer properties, then under the causal theory of properties one will not have distinct realizer properties and one hence lacks distinct realizations. But if one has distinct realizer properties then these properties must plausibly play different causal roles defining their respective realized property instances and under the causal theory the realized properties are thus distinct, again precluding multiple realization of a common property.[13]

Defenders of the dimensioned view ought to object immediately to the dilemma that they can make perfectly good sense of a multiple realization of a property *T* where the "causally relevant properties" differ. How? They can do this by virtue of the dimensioned view allowing realizers to be properties/relations of an individual's constituents. Although these properties/relations of constituents differ in the powers that they contribute, we nonetheless have distinct realizations of *T* because these different properties/relations both result in the causal powers individuative of *T*. This implication of the dimensioned metaphysics renders the first horn of the dilemma blunt and is arguably built into the received view of laws in the special sciences.[14] For on the received view, it is argued that these laws often range over special science properties whose instances are realized by a heteroge-

[13] One may be concerned about the validity of such reasoning even under the flat view. For example, in his "Realization and Mental Causation," Shoemaker notes that it is possible for distinct properties to each contribute the powers individuative of some realized property, and in some sense play its causal role, while these properties differ in their other contributions of powers. Given the latter possibility, even under the flat metaphysics, it does not follow that because two property instances contribute the powers individuative of some realized property that there cannot be an instance of multiple realization. For these properties can still be distinct in virtue of differences in their *other* contributions of powers. This provides a potential route to blunt the second horn of Shapiro's dilemma even under the flat view and similar points also hold under the dimensioned account.

[14] For example, in Fodor's "Special Sciences."

neous array of distinct realizer properties/relations instantiated in the constituents of individuals having the special science properties.

Similar points loom large for another recent critique of the received view. Clapp has argued at length that disjunctions of predicates may refer to a property if all of the predicates refer to, or "overlap on," a common set of causal powers (*op. cit.*, pp. 124–32). Clapp also explicitly assumes the flat view of realization (*op. cit.*, p. 129) which implies each realizer of some property *H* ***must*** contribute a common set of powers in the powers that individuate *H*. Building on these claims, Clapp argues that a common property will consequently be referred to by a disjunction of realizer predicates, that is, a disjunction of predicates each of which refers to a realizer of a particular realized property. And given this conclusion, Clapp rejects (*op. cit.*, p. 132) a central premise of the so-called "argument from multiple realizability" that has been used by proponents of the received view, such as Fodor, against classical, Nagelian reduction. For the premise in question claims that realizer properties will often be heterogeneous and that disjunctions of realizer predicates will very often *not* refer to a common property (*op. cit.*, p. 122). Taking himself to have undermined one of the argument from multiple realizability's premises, Clapp also presses a revisionary view of the special sciences.

Once again, we can see that by themselves Clapp's criticisms do not suffice for a legitimate objection to the received view or its weapon in the argument from multiple realizability. For as we noted in considering Shapiro's dilemma, under the dimensioned metaphysics it is simply not true that realizer properties must *contribute* a common set of powers. Obviously under the dimensioned view, all realizers *result in* the powers of the realized property, but they may, and often do, contribute no common powers, including those individuative of the realized property. Consequently, the proponent of the received view will deny Clapp's key assumption in his argument to the conclusion that disjunctions of realizer predicates must refer to a common property. For the defender of the received view will not accept that realizer properties must all be individuated by a common set of powers. Furthermore, the defender of the received view will presumably reiterate the empirical evidence, based upon the realizers often found in constituents, that they have long argued supports the heterogeneity of realizer properties and the consequent implication that realizers often contribute no common causal powers.

The latter response nicely highlights what we have found about the role of the dimensioned account of realization as a keystone of the integrated package of claims that constitute the received view. The dimensioned account plausibly underpins the received view of the wide

extent of MR, the heterogeneity of realizers, and, consequently, its distinctive picture of special sciences and their laws. Like Shapiro, and again as a result of not distinguishing the different metaphysical views of realization, Clapp has failed to appreciate the central role that the dimensioned metaphysics plays in the received view. Consequently, Clapp also fails to see that his assumption of the flat account undermines the dialectical effectiveness of his putative objections.

In concluding, I should again emphasize that many issues about multiple realizability, and realization itself, remain unresolved. One might therefore try to salvage Shapiro or Clapp's critiques by providing reasons to accept the flat picture of realization instead of the dimensioned perspective. But I hold out little hope for such an endeavor. For I have argued elsewhere that there are good reasons to reject the flat view in favor of the dimensioned account, since only the latter covers all the cases of realization illuminated by the sciences.[15] More importantly, and whatever the success of the latter arguments may be, I have shown here that Shapiro has been too quick in his criticisms about MR. Furthermore, I have suggested that Shapiro and Clapp's arguments for revisionary views of the special sciences and their laws are also ineffective. For we have now found that the debates over the natures of realization, MR and the special sciences are interconnected. And Shapiro and Clapp both assume the flat metaphysics in their criticisms and defenders of received view will use their dimensioned account of realization to deny key premises of these critiques.

My final conclusion is therefore that Shapiro's, and Clapp's, recent criticisms of the received view of MR and the special sciences beg the question or are otherwise ineffective, for their conclusions do not go through under the dimensioned account of realization around which the received view is arguably founded. Looking more widely, I suggest that these failed critiques show that philosophers of psychology must pay careful attention to the metaphysics of realization and can only benefit by joining the ongoing project of illuminating the natures of both the realization relation and multiple realizability.

CARL GILLETT

Illinois Wesleyan University

[15] See my "The Dimensions of Realization," and also "Moving beyond the Subset Model of Realization."

NEW BOOKS: PAPERBACK REPRINTS

CLARK, PETER, and KATHERINE HAWLEY, eds.: *Philosophy of Science Today.* New York: Oxford, 2000/2003. 312 p. Paper $19.95.

DRETSKE, FRED I.: *Knowledge and the Flow of Information.* Stanford: CSLI, 1981/2003. xiv, 273 p. Paper $23.95.

FISHER, PHILIP: *Wonder, the Rainbow, and the Aesthetics of Rare Experiences.* Cambridge: Harvard, 1998/2003. 191 p. Paper $17.95.

FOOT, PHILIPPA: *Virtues and Vices.* New York: Oxford, 1978/2002. xvi, 213 p. Paper $19.95.

GARBER, MARJORIE: *Academic Instincts.* Princeton: University Press, 2001/2003. xi, 187 p. Paper $14.95.

HEIDEGGER, MARTIN: *Plato's Sophist.* RICHARD ROJCEWICZ and ANDRÉ SCHUWER, trans. Bloomington: Indiana UP, 1997/2003. xxvii, 476 p. Paper $29.95.

KANT, IMMANUEL: *Critique of Pure Reason.* NORMAN KEMP SMITH, trans. New York: Palgrave MacMillan, 1929/2003. xxxi, 681 p. Paper $22.95.

KUKLICK, BRUCE: *A History of Philosophy in America, 1720–2000.* New York: Oxford, 2002/2003. xiii, 328 p. Paper $19.95.

MELE, ALFRED R.: *Self-Deception Unmasked.* Princeton: University Press, 2000/2001. xiii, 98 p. Paper $15.95.

MIDGLEY, MARY: *Science and Poetry.* New York: Routledge, 2001/2002. x, 230 p. Paper $14.95.

MOHANTY, J. N.: *The Self and Its Other: Philosophical Essays.* New York: Oxford, 2002/2003. xv, 146 p. Paper $13.95.

NEUHOUSER, FREDERICK: *Foundations of Hegel's Social Theory.* Cambridge: Harvard, 2000/2003. xiii, 337 p. Paper $29.95.

NOZICK, ROBERT: *Invariances: The Structure of the Objective World.* Cambridge: Harvard, 2001/2003. x, 416 p. Paper $19.95.

PAREKH, BHIKHU: *Rethinking Multiculturalism: Cultural Diversity and Political Theory.* Cambridge: Harvard, 2000/2002. xii, 379 p. Paper $17.95.

POPPER, KARL: *Conjectures and Refutations: The Growth of Scientific Knowledge.* New York: Routledge, 1963/2002. xvi, 582 p. Paper $31.95; *The Logic of Scientific Discovery.* New York: Routledge, 1992/2002. xxvii, 513 p. Paper $14.95; *The Poverty of Historicism.* New York: Routledge, 1957/2002. xiii, 156 p. Paper $22.95; *Unended Quest. An Intellectual Autobiography.* New York: Routledge, 1992/2002. x, 315 p. Paper $32.95.

RORTY, RICHARD, DEREK NYSTROM, and KENT PUCKETT: *Against Bosses, Against Oligarchies.* Chicago: Prickly Paradigm, 1998/2002. xii, 66 p. Paper $10.00.

ROSEN, STANLEY: *The Ancients and the Moderns: Rethinking Modernity.* South Bend, IN: St. Augustine's, 1989/2002. x, 236 p. Paper $23.00; *The Question of Being. A Reversal of Heidegger.* South Bend, IN: St. Augustine's, 1993/2002. xxiii, 342 p. Paper $27.00.

ROSENZWEIG, FRANZ: *On Jewish Learning.* N. N. GLATZER, ed. Madison: Wisconsin UP, 1955/2002. 128 p. Paper $16.95.

ROSS, DAVID: *The Right and the Good.* PHILIP STRATTON-LAKE, ed. New York: Oxford, 1930/2002. lviii, 183 p. Paper $18.95.

OUTSTANDING SCHOLARSHIP FROM CAMBRIDGE

The Cambridge History of Philosophy 1870-1945
Edited by Thomas Baldwin
0-521-58104-X Hardback $120.00*

The Innovations of Idealism
Rudiger Bubner
Modern European Philosophy
0-521-46262-1 Hardback $55.00

The Evolution of Reason
Logic as a Branch of Biology
William S. Cooper
Cambridge Studies in Philosophy and Biology
0-521-54025-9 Paperback $28.00*

The Science of the Struggle for Existence
On the Foundations of Ecology
Gregory J. Cooper
Cambridge Studies in Philosophy and Biology
0-521-80432-9 Hardback $60.00

An Introduction to the Philosophy of Art
Richard Eldridge
0-521-80135-4 Hardback $65.00
0-521-80521-X Paperback $23.00

The Cambridge Companion to Medieval Jewish Philosophy
Edited by Daniel H. Frank and Oliver Leaman
Cambridge Companions to Philosophy
0-521-65207-3 Hardback $75.00
0-521-65574-9 Paperback $27.00

Richard Rorty
Edited by Charles Guignon and David R. Hiley
Contemporary Philosophy in Focus
0-521-80058-7 Hardback $55.00
0-521-80469-2 Paperback $20.00

The Political Philosophy of Needs
Lawrence Hamilton
0-521-82782-5 Hardback $60.00

Iconoclasm in Aesthetics
Michael Kelly
0-521-82209-2 Hardback $60.00

Novalis: *Fichte Studies*
Edited by Jane Kneller
Cambridge Texts in the History of Philosophy
0-521-64353-8 Hardback $58.00
0-521-64392-9 Paperback $22.00

The Value of Knowledge and the Pursuit of Understanding
Jonathan L. Kvanvig
Cambridge Studies in Philosophy
0-521-82713-2 Hardback $60.00

The Cambridge Companion to Medieval Philosophy
Edited by A.S. McGrade
Cambridge Companions to Philosophy
0-521-80603-8 Hardback $60.00
0-521-00063-7 Paperback $24.00

A Physicalist Manifesto
Thoroughly Modern Materialism
Andrew Melnyk
Cambridge Studies in Philosophy
0-521-82711-6 Hardback $70.00

Faking It
William Ian Miller
0-521-83018-4 Hardback $29.00

Biological Complexity and Integrative Pluralism
Sandra D. Mitchell
Cambridge Studies in Philosophy and Biology
0-521-81753-6 Hardback $70.00
0-521-52079-7 Paperback $26.00

John Searle
Edited by Barry Smith
Contemporary Philosophy in Focus
0-521-79288-6 Hardback $65.00
0-521-79704-7 Paperback $20.00

Kierkegaard's Relations to Hegel Reconsidered
Jon Stewart
Modern European Philosophy
0-521-82638-4 Hardback $55.00

Religion after Metaphysics
Edited by Mark Wrathall
0-521-82498-2 Hardback $55.00*
0-521-53196-9 Paperback $20.00*

*Prices subject to change



CAMBRIDGE UNIVERSITY PRESS 800-872-7423 us.cambridge.org/philosophy

Prometheus Books

BEHOLD THE ANTICHRIST
Bentham on Religion
Delos B. McKown

During his long, productive life the eminent English philosopher and exponent of utilitarianism Jeremy Bentham (1748–1832) wrote not just on political philosophy but also clandestinely on religion. Under the pseudonym of Philip Beauchamp he indirectly published an attack on natural religion called *Analysis of the Influence of Natural Religion on the Temporal Happiness of Mankind* and under the pseudonym of Gamaliel Smith he brought out (also indirectly) a book of New Testament criticism called *Not Paul, But Jesus*. Earlier, he had bravely released under his own name *Church-of-Englandism and Its Catechism Examined*. Philosopher Delos B. McKown introduces these major treatises on religion, and then presents an extensive synopsis of each. This thorough analysis of three little-known works by one of philosophy's great minds makes an outstanding contribution to Bentham scholarship and will be of interest to humanists and philosophers of religion.
365 pp • ISBN 1-59102-116-2 • HC $32

Humanity Books
an imprint of Prometheus Books

SKEPTICISM IN RENAISSANCE AND POST-RENAISSANCE THOUGHT
New Interpretations
Edited by José R. Maia Neto and Richard H. Popkin

This second volume in the *Journal of the History of Philosophy* book series is devoted to the resurgence of skepticism in the Renaissance and after. It contains eight original essays by historians of early modern philosophy, with concluding remarks by Richard H. Popkin, who reviews fifty years of scholarship on the history of early modern skepticism and evaluates its present stage.
200 pp • ISBN 1-59102-156-1 • HC $55

KROPOTKIN
The Politics of Community
Brian Morris

Morris seeks to affirm the contemporary relevance of Kropotkin as a political and moral philosopher, and as a social ecologist. An important contribution to the history of anarchism and to contemporary debates in political theory and social ecology.
340 pp • ISBN 1-59102-158-8 • HC $50

HEGEL'S *PHENOMENOLOGY OF SPIRIT*
New Critical Essays
Edited by Alfred Denker and Michael Vater

Offers an interpretation of Hegel's *Phenomenology* as well as a concise reading of the main text. The essays, written by both established Hegel scholars and promising young researchers, present the reader with an international overview of recent Hegel research.
285 pp • ISBN 1-59102-056-5 • PB $35

For exam copies contact LRislo@prometheusbooks.com
Call toll free (800) 421-0351 • Fax (716) 691-0137 • www.prometheusbooks.com

Philosophy

CSLI PUBLICATIONS

New from the Center for the Study of Language and Information

Exploring Internal Tensions in Wittgenstein's *Tractatus*

An Analysis of the Conditions for the Possibility of Language

María Cerezo

In this volume, María Cerezo examines Wittgenstein's *Tractatus* as a response to some of Frege's and Russel's logical problems. In analyzing the tractarian conditions for the possibility of language, she explains the two main theories of the proposition in *Tractatus*: the truth-functions theory and the picture theory.

Paper $25.00

Meaning, Basic Self-Knowledge, and Mind

Essays on Tyler Burge

Edited by María Frápolli and Esther Romero

This volume comprises a lively and thorough discussion between philosophers and Tyler Burge about Burge's recent, and already widely accepted, position on the theory of meaning, mind, and knowledge.

Paper $25.00

Reference and the Rational Mind

Kenneth A. Taylor

To demonstrate his view of referentialism, Kenneth A. Taylor offers original and provocative accounts of a wide variety of semantic, pragmatic, and psychological phenomena, such as empty names, propositional attitude contexts, the nature of concepts, and the ultimate source and nature of normativity

Paper $25.00

The Philosopher's Annual, Volume 24

Edited by Patrick Grim, Peter Ludlow, and Gary Mar

This volume presents the ten best articles in philosophy of 2001, including papers by Robert Bernasconi, Hans Halvorson, Christopher Hitchcock, Ignacio Jane, Brian Leiter, Liam Murphy and Thomas Nagel, Joel Pust, Alison Simmons, Jason Stanley and Timothy Williamson, and Crispin Wright.

Paper $30.00

Narrative Theory and the Cognitive Sciences

Edited by David Herman

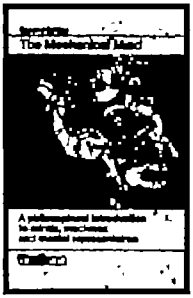
Research on human intelligence has postulated that studying narrative communication can provide important insight into the roots of self. In that spirit, this volume focuses on narrative as a crossroads where cognition, psychology, linguistics, literary theory, and cognitive narratology meet.

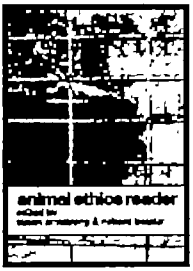
Paper $27.50

Distributed by

The University of Chicago Press

1427 East 60th Street, Chicago, IL 60637 www.press.uchicago.edu









Philosophy from Routledge

THE AFRICAN PHILOSOPHY READER

Second Edition

Edited by P.H. Coetzee and A.P.J. Roux

This pathbreaking work reveals the way in which traditional philosophical issues related to ethics, metaphysics, and epistemology take on specific forms in Africa's postcolonial struggles

672 pp • $32.95

THE ANIMAL ETHICS READER

Edited by Susan Armstrong and Richard Botzler

"A comprehensive, up-to-date anthology going beyond the usual animal rights issues should be of great use, not just to teachers, but to the growing non-professionals interested in these matters."

—Neil Shanks, Tennessee State University

608 pp • $29.95

ROUTLEDGE PHILOSOPHY GUIDEBOOK TO HUSSERL AND THE *CARTESIAN MEDITATIONS*

A.D. Smith

This book is an introduction to Husserl's phenomenology and is based on Descartes' *Meditations on First Philosophy*.

296 pp • $18.95

CLASSICAL PHILOSOPHY: A CONTEMPORARY INTRODUCTION

Christopher Shields

"This is a fine introduction to ancient philosophy, consistently clear and carefully argued, interesting and generally attractive to read."

—Vasilis Politis, Trinity College, Dublin

176 pp • $20.95

LIVING PHILOSOPHY

An Introduction to Moral Thought

Third Edition

Ray Billington

"This is philosophy that doesn't hide; it demonstrates powerfully how the philosophical stance can illuminate things that matter to many people. Direct, argumentative, informative."

—John Shand, The Open University

376 pp • $27.95

MECHANICAL MIND

A Philosophical Introduction to Minds, Machines and Mental Representation

Second Edition

Tim Crane

"Terrifically useful...It gives the best simple introduction that I know to the computational theory of mind."

—John Searle, University of California, Berkeley

272 pp • 11 line drawings • $16.95

Routledge

1.800.634.7064 □ WWW.ROUTLEDGE-NY.COM

Profound.

Philosophical Analysis in the Twentieth Century

Scott Soames

This is a major, wide-ranging history of analytic philosophy since 1900, told by one of the tradition's leading contemporary figures. The first volume takes the story from 1900 to mid-century. The second brings the history up to date.

Scott Soames has been at the center of some of the tradition's most important debates and he writes with exceptional ease about its often complex ideas.

"A history of analytic philosophy and an excellent piece of analytic philosophy in its own right."
—Kit Fine, New York University

Volume 1: The Dawn of Analysis
Cloth $35.00 ISBN 0-691-11573-7
Volume 2: *The Age of Meaning*
Cloth $35.00 ISBN 0-691-11574-5 Due December

A Defense of Hume on Miracles

Robert J. Fogelin

Since its publication in the mid–eighteenth century, Hume's discussion of miracles has been the target of severe and often ill-tempered attacks. In this book, one of our leading historians of philosophy offers a systematic response to these attacks.

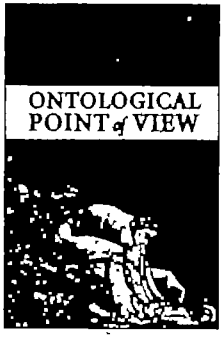
Robert Fogelin responds first to some early critics of Hume's argument and then to two recent critics, David Johnson and John Earman. Fogelin's goal is to show that Hume's treatment of miracles has a coherence, depth, and power that helps it remain the best work on the subject.

Princeton Monographs in Philosophy Harry Frankfurt, series editor
Cloth $24.95 ISBN 0-691-11430-7 Due October

PRINCETON University Press

800-777-4726 • READ EXCERPTS ONLINE
WWW.PUP.PRINCETON.EDU

New from
Oxford







How We Act
Causes, Reasons, and Intentions
BERENT ENC
Berent Enc provides a philosopher's sustained examination of the distinctions between voluntary and non-voluntary actions: he portrays action as belonging to the causal order of events in nature, a theory from which new and surprising accounts of intention and voluntary action emerge.
November 2003 $39 95

From an Ontological Point of View
Edited by JOHN HEIL
From an Ontological Point of View is a highly original and accessible exploration of fundamental questions about what there is. John Heil discusses such issues as whether the world includes levels of reality; what makes things true; qualities, powers, and the relation these bear to one another. He advances an account of the fundamental constituents of the world around us
2003 $45 00

Conceptions of Truth
WOLFGANG KÜNNE
Truth is one of the most debated topics in philosophy; Wolfgang Kunne presents a comprehensive critical examination of all major theories, from Aristotle to the present day. The clarity of exposition and the wealth of examples will make *Conceptions of Truth* an invaluable and stimulating guide for advanced students and scholars.
2003 $55 00

Mindreading
SHAUN NICHOLS and STEPHEN P. STICH
The everyday capacity to understand the mind plays an enormous role in our ordinary lives. Nichols and Stich provide a detailed and integrated account of the intricate web of mental components underlying this fascinating and multifarious skill.
Mindreading is another trailblazing volume in the prestigious interdisciplinary *Oxford Cognitive Science Series*
November 2003 paper $24 95 cloth $74 00

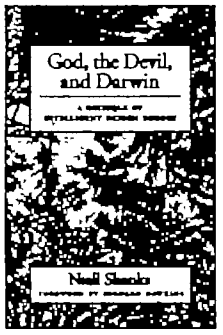
Effortless Action
Wu-wei as Conceptual Metaphor and Spiritual Ideal in Early China
EDWARD SLINGERLAND
"Edward Slingerland is one of a group of exciting and creative young scholars revolutionizing the study of Chinese history, culture, and religion.... *Effortless Action* is a remarkable work that explores the meaning of the crucial concept of wu-wei in a depth never before achievable, showing how Chinese metaphorical thought forms a nexus around this most central of ideas."—George Lakoff, Professor of Linguistics, *University of California, Berkeley.*
2003 $60 00

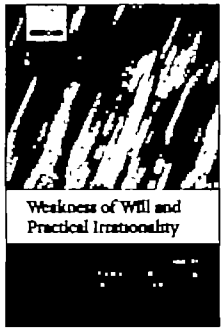
Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556. Visit our website at www.oup-usa.org

OXFORD

New from
Oxford







Desert and Justice

Edited by SERENA OLSARETTI

Does justice require that individuals get what they deserve? Serena Olsaretti brings together new essays by leading moral and political philosophers examining the relation between desert and justice; they also illuminate the nature of distributive justice, and the relationship between desert and other values, such as equality and responsibility.

(Mind Association Occasional Series)

2003 $65 00

Action and Its Explanation

DAVID-HILLEL RUBEN

David-Hillel Ruben mounts a defense of some unusual and original positions in the philosophy of action. Written from a point of view out of sympathy with the assumptions of much of contemporary philosophical action theory, his book draws its inspiration from philosophers as diverse as Aristotle, Berkeley, and Marx.

2003 $45 00

God, the Devil, and Darwin

A Critique of Intelligent Design Theory

NIALL SHANKS

Foreword by RICHARD DAWKINS

Niall Shanks has written the first accessible introduction to, and critique of, *Intelligent Design theory* (ID): the contention that life is so complex that the only plausible explanation is the existence of an intelligent designer. Shanks argues that it is in fact a serious threat to the scientific and democratic values that are our cultural and intellectual inheritance from the Enlightenment.

November 2003 $29 95

Weakness of Will and Practical Irrationality

Edited by SARAH STROUD and CHRISTINE TAPPOLET

Sarah Stroud and Christine Tappolet present eleven original essays on weakness of will, a topic straddling the divide between moral philosophy and philosophy of mind, and the subject of much current attention. An international team of established scholars and younger talent provide perspectives on all the key issues in this fascinating debate, the book will be essential reading for anyone working in the area.

November 2003 $45 00

Hume Variations

JERRY A FODOR

Jerry Fodor looks to David Hume for help in advancing our understanding of the mind. Going back to his work after more than 250 years we find that Hume is remarkably perceptive about the components and structure that a theory of mind requires. Careful study of his *Treatise of Human Nature* helps us to see what is amiss with much twentieth-century philosophy of mind, and to get on the right track.

November 2003 $22 00

Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556. Visit our website at www.oup-usa.org

OXFORD



Saving the Differences

Essays on Themes from *Truth and Objectivity*

CRISPIN WRIGHT

Crispin Wright's *Truth and Objectivity* brought about a far-reaching reorientation of the metaphysical debates concerning realism and truth. The essays in this companion volume prefigure, elaborate, or defend the proposals put forward in that landmark work.

"*Truth and Objectivity* is a strikingly resourceful and serious book, imbued with respect for the difficulty of philosophical problems and a readiness to probe them with all the conceptual instruments of contemporary analytic philosophy."
—Timothy Williamson, *International Journal of Philosophical Studies*

New in cloth

Justice, Luck, and Knowledge

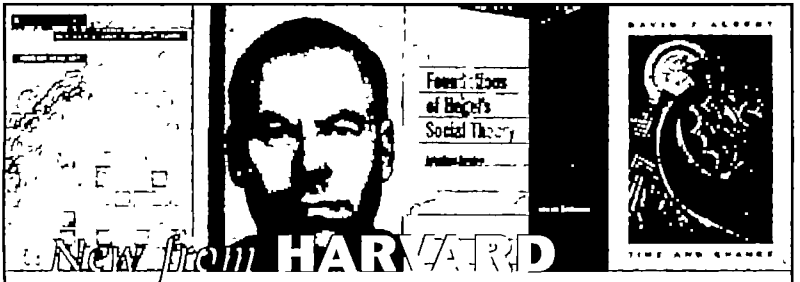
S. L. HURLEY

"[A] rich, impressive book that is required reading by researchers in philosophy and cognitive science."
—Joshua Weinberg, *Journal of Consciousness Studies*

"*Consciousness in Action*... exhibits the astonishing breadth of knowledge, technical virtuosity and subtle analyses Hurley's readers have come to expect...[its] integration of philosophy with neuropsychology and cognitive science provides new avenues of research for philosophers concerned about the nature of the mind, perception, and action...[The] book's impact will continue to be felt for years to come."
—Dan Silber, *Philosophy in Review*

New in cloth

HARVARD UNIVERSITY PRESS 800.405.1619 www.hup.harvard.edu



Bigger than Chaos

Understanding Complexity through Probability

MICHAEL STREVENS

"This impressive book tackles an important question: how can systems of many interacting parts, which thus display low-level complexity, give rise to high-level simplicity?... [Strevens] makes real progress on a genuinely difficult topic, one that is of central interest to science and to the philosophy of science. He also has a seemingly effortless command of his materials and a sure grip on the conceptual issues. The work is technically sophisticated...and elegantly written. This is what good philosophy is all about."

—Alan Hajek,
California Institute of Technology

New in cloth

NEW IN PAPERBACK

Time and Chance

DAVID Z ALBERT

"As usual with Albert's work, the exposition is brisk and to the point, and exceptionally clear...The book will be an extremely valuable contribution to the literature on the subject of philosophical issues in thermodynamics and statistical mechanics."

—Lawrence Sklar,
University of Michigan

Foundations of Hegel's Social Theory

Actualizing Freedom

FREDERICK NEUHOUSER

"This is a fine book, and it will be a significant contribution both to Hegel scholarship and to contemporary philosophical discussions of modern ethical life...There is no question in my mind that it will play a long and decisive role in future discussions of Hegel."

—Robert B. Pippin,
University of Chicago

HARVARD UNIVERSITY PRESS 800.405.1619 www.hup.harvard.edu

THE JOURNAL OF PHILOSOPHY

Celebrates 100 Years

Special Limited Time Offer

Full Volumes 1-99

1904-2002

$50.00 per Volume

All volumes are available for $50.00 plus $5.00 shipping and handling.

Send check or money order to:

The Journal of Philosophy

1150 Amsterdam Avenue, MC 4972

Columbia University

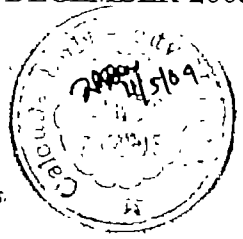
New York, NY 10027

(212) 666-4419

www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

VOLUME C, NUMBER 12
DECEMBER 2003



page

605 *The Role of Variables.* Kit Fine

BOOK REVIEWS

632 Daniel C. Dennett: *Freedom Evolves* John Martin Fischer

638 Jeremy Waldron: *God, Locke, and Equality: Christian Foundations in Locke's Political Thought* Partha Chatterjee

642 NOTES AND NEWS

644 NEW BOOKS

651 INDEX TO VOLUME C

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX: (212) 932-8721.

You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2003

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset and printed by Capital City Press, Montpelier, VT.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

POSTMASTER: Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027

© Copyright 2003 by the Journal of Philosophy, Inc.

ISSN 0022-362X

,826

THE JOURNAL OF PHILOSOPHY
VOLUME C, NO. 12, DECEMBER 2003

THE ROLE OF VARIABLES*

It is generally supposed—by logicians and philosophers alike—that we now possess a perfectly good understanding of how variables work in the symbolism of logic and mathematics. Once Gottlob Frege had provided a clear syntactic account of variables and once Alfred Tarski had supplemented this with a rigorous semantic account, it would appear that there was nothing more of any significance to be said. It seems to me, however, that this common view is mistaken. There are deep problems concerning the role of variables that have never been properly recognized, let alone solved, and once we attempt to solve them we see that they have profound implications not only for our understanding of variables but also for our understanding of other forms of expression and for the general nature of semantics.

It is my aim here to say what these problems are, to explain how they are to be solved, and to indicate the implications for the rest of semantics. I begin with an antinomy concerning the role of variables which I believe any satisfactory account of them should solve (section I). I then argue that the three main semantical schemes currently on the market—the Tarskian, the instantial, and the algebraic—are unsuccessful in solving the puzzle (sections II–III) or in providing a satisfactory semantics for first-order logic (sections IV–V). Finally, I offer an alternative scheme that is capable of solving the antinomy (section VI) and of providing a more satisfactory semantics for first-order logic (section VII). It is based upon a new approach to representational semantics, which I call *semantic relationism*, and I conclude by

* I should like to thank the members of seminars at St Andrews University, New York University and University of California/Los Angeles for helpful discussion. I am especially grateful to Ruth Chang and Louis Deroset for their comments on earlier drafts.

0022-362X/03/0012/605–31 © 2003 The Journal of Philosophy, Inc.

considering the implications of this approach for other topics in linguistics and the philosophy of language (section VIII).

I. THE ANTINOMY OF THE VARIABLE

At the root of the present considerations is a certain puzzle. In order to state this puzzle, we shall need to appeal to the notion of *semantic role*. I do not mean this as a technical notion of a kind that one might find in a formal semantics, but as a nontechnical notion whose application may already be taken to be implicit in our intuitive understanding of a given language or symbolism. For in any meaningful expression, there is something conventional about the expression—having to do with the actual symbols or words used, and something nonconventional—having to do with the linguistic function of those symbols or words. And 'semantic role' is just my term for this linguistic or nonconventional aspect of a meaningful expression.

In fact, there will be no need, in stating the puzzle, to appeal to semantic role per se. For our interest will solely be in whether two expressions have the *same* or a *different* semantic role. And, of course, their having the same, or a different, semantic role is compatible with different views as to what that semantic role might be. Indeed, our puzzle would appear to arise under any reasonable conception of semantic role—be it Russellian, Fregean, or of some other kind.

Now for the puzzle. Suppose that we have two variables, say 'x' and 'y'; and suppose that they range over the same domain of individuals, say the domain of all reals. Then it appears as if we wish to say contradictory things about their semantic role. For when we consider their semantic role in two distinct expressions—such as '$x > 0$' and '$y > 0$', we wish to say that their semantic role is the same. Indeed, this would appear to be as clear a case as any of a mere "conventional" or "notational" difference; the difference is merely in the choice of the symbol and not in its linguistic function. On the one hand, when we consider the semantic role of the variables in the same expression—such as '$x > y$'—then it seems equally clear that their semantic role is different. Indeed, it is essential to the linguistic function of the expression as a whole that it contains two distinct variables, not two occurrences of the same variable, and presumably this is because the roles of the distinct variables are not the same.

Generalizing from our example, we arrive at the following two claims:

Semantic Sameness (SS): Any two variables (ranging over a given domain of objects) have the same semantic role; and

Semantic Difference (SD): Any two variables (ranging over a given domain of objects) have a different semantic role.

Yet we cannot univocally maintain both (at least given that there are at least two variables!).

We might call this puzzle the *antinomy of the variable*. It was first stated by Bertrand Russell, though in ontological rather than semantical fashion. He writes, "x is, in some sense, the object denoted by *any term*; yet this can hardly be strictly maintained, for different variables may occur in a proposition, yet the object denoted by *any term*, one would suppose, is unique."[1] It also bears a close affinity with Frege's puzzle concerning names,[2] though with variables taking the place of names.

Clearly, any resolution of the conflict between (SS) and (SD) must begin by locating an ambiguity in the phrase 'same semantic role'. Whatever the sense in which the variables x and y are semantically the same in '$x > 0$' and '$y > 0$' cannot be the sense in which they fail to be semantically the same in '$x > y$'. Now as a first stab towards locating the ambiguity, we might appeal to a notion of context. The variable 'x' in the context of the formula '$x > 0$' plays the same semantic role as the variable 'y' in the context of the formula '$y > 0$'. On the other hand, the two variables x and y will play different semantic roles within the context of the single formula '$x > y$'. Thus (SS) will hold for sameness of semantic role in the *cross*contextual sense while (SD) will hold for difference of semantic role in the *intra*contextual sense; and, given the ambiguity in the respective senses of "same" and "different" role, contradiction may be avoided.

Natural as this response may be, it does not really solve the puzzle but merely pushes its solution back a step. For why do we say that the variables x and y have a different semantic role in '$x > y$'? Clearly, it has to do with the fact that the occurrence of y cannot be replaced by x without a change in semantic role; the role of the pair of distinct variables x, y in '$x < y$' is different from the role of the pair of the identical variables x, x in '$x < x$'. In other words, the intracontextual difference in semantic role between the variables x and y within the single formula '$x > y$' amounts to a crosscontextual difference in semantic role between the pair of variables x, y in '$x > y$' and the pair of variables x, x in '$x > x$'. And, in general, to say that there is an intracontextual difference between x and y, in the intended sense, is just to say that there is crosscontextual difference between the pairs of variables x, y and x, x.

[1] In §93 of *Principles of Mathematics* (London: Allen and Unwin, 1903).

[2] Stated at the beginning of "Uber Sinn und Bedeutung," republished as "On Sense and Reference" in *Translations from the Philosophical Writings of Gottlob Frege*, Peter T. Geach and Max Black, eds. (New York: Oxford, 1952).

We may therefore state (SS) and (SD) in the following form:

(SS)′: There is no crosscontextual difference in semantic role between the variables x and y, and
(SD)′: There is a crosscontextual difference in semantic role between the pair of variables x, y and the pair x, x,

using an univocal notion of semantic role throughout.

In contrast to the earlier formulation, there is now no explicit contradiction. But there is still a difficulty in seeing how the two claims, (SS)′ and (SD)′, might be reconciled. For how can there be a crosscontextual difference in semantic role between the pair of variables x, y and the pair x, x unless there is a crosscontextual semantic difference between the variables x and y themselves? What else could account for the difference in semantic role between the pairs x, y and x, x except a semantic difference in the individual variables x and y? Or to put it another way, if there is a semantic difference between x, y and x, x, then there must be a semantic difference between x and y, and it is hard to see why this difference should only be "turned on" or made manifest when the variables are used in the same context and not when they are used in different contexts.

The puzzle therefore remains; and any solution to the puzzle should either explain how (SS)′ and (SD)′ might be compatible, notwithstanding appearances to the contrary, or it must explain how one of (SS)′ or (SD)′ might reasonably be rejected.

II. THE TARSKIAN APPROACH

It might be thought that the solution to our puzzle should be sought in the various semantics that have been developed for the language of predicate logic. After all, it is presumably the aim of these semantics to account for the semantic role of the expressions with which they deal; and so we should expect them to account, in particular, for the semantic role of variables.

When we turn to the various semantics that have in fact been developed, however, we find them entirely unsuited to the purpose. Let us begin with the most familiar of them, that of Tarski.[3] The reader will recall that the Tarski semantics proceeds by characterizing

[3] In "Der Wahrheitsbegriff in den Formalisierten Sprachen," *Studia Philosophica*, 1 (1933): 261–405, and republished in translation in *Logic, Semantics, and Metamathematics* (New York: Oxford, 1955), pp. 152–278. There are other approaches to the semantics of first-order logic—for example, my own approach in terms of arbitrary objects in *Reasoning with Arbitrary Objects* (New York: Blackwell, 1985)—that I have not considered. But I do not believe that there are any more successful in dealing with the antinomy.

a relation of satisfaction between assignments and formulas. To fix our ideas, let us suppose that the variables of our language are x_1, x_2,... and that the domain of discourse is $\mathcal{D}$. We may then take an *assignment* to be a function taking each variable of the language into an individual from $\mathcal{D}$; and the semantics will specify—by means either of a definition or of a set of axioms—what it is for each kind of formula to be satisfied by an assignment. It will state for example that an assignment satisfies the formula $\sim B$ just in case it fails to satisfy B or that an assignment satisfies the formula $\forall x B$ just in case every x-variant of the assignment (differing at most on the value assigned to x) satisfies B.

Now what account, within the framework of the Tarski semantics, can be given of the semantic role of the variables? There would appear to be only two options. The first is to take the semantic role of a variable to be given by its range of values (the domain $\mathcal{D}$ in the case above). Indeed, quite apart from the connection with the Tarski semantics, this is the usual way of indicating how a variable is to be interpreted: one specifies its range of values; and that is it.

This approach does indeed account for the fact that the semantic role of any two variables x and y (with an identical range of values) is the same. But it does nothing to account for the semantic difference between the pairs of variables x, y and x, x; and nor is any reason given for denying that there is an intuitive difference in semantic role.

The other option is to take the semantic role of a variable to be what one might call its *semantic value* under the given semantics. The semantic values are those entities which are assigned (or which might be taken to be assigned) to the meaningful expressions of the language and with respect to which the semantics for the language is compositional. When we examine the Tarski semantics, we see that the semantic value of an open formula (one possibly containing free variables) might be taken to be the function that takes each assignment into the "truth value" of the formula under that assignment (the TRUE if the formula is satisfied by the assignment and the FALSE otherwise) and, similarly, the semantic value of an open term might be taken to be the function that takes each assignment into the denotation of the term under that assignment. Thus the semantic value of the formula '$x > 0$' (under the natural interpretation of the language) would be the function that takes any assignment into TRUE if the number it assigns to x is positive and into FALSE otherwise; and the semantic value of the term '$x + y$' would be the function which takes an assignment into the sum of the numbers which it assigns to x and y. We then easily see that the Tarski semantics is compositional with respect to the semantic values as so conceived; it

"computes" the semantic value of a complex expression on the basis of the semantic values of the simpler expressions from which it is derived.

Under this approach, the semantic value of a variable x will be a special case of the semantic value of a term; it will be a function which takes each assignment into the individual which it assigns to x. It is therefore clear, if we identify semantic roles and semantic values, that x and y will differ in their semantic roles; for if we take any assignment which assigns different individuals to x and y (this requires, of course, that there be at least two individuals in the domain), then the semantic value of x will deliver the one individual in application to that assignment while the semantic value of y will deliver the other individual in application to the assignment.

We therefore secure a semantic difference between the pairs x, y and x, x under this account of semantic role. But we are at a loss to account for the fact that the semantic role of the variables x and y is the same; and nor is any reason given for disputing the intuitive identity in semantic role.

There is another, perhaps even more serious, problem with the approach. For although it posits a difference between x and y (and hence between the pairs x, y and x, x), it does nothing to account for the *semantic* difference. For in the last analysis, the posited difference between the semantic values for x and y simply turns on the difference between the variables x and y themselves. The semantic value of the variable x, for example, delivers one individual rather than another in application to a given assignment simply because it is the variable that it is. Thus what we secure, strictly speaking, is not a *semantic* difference, one lying on the nonconventional side of language, but a *typographic* difference, one lying purely on the conventional side of language; and so we have done nothing to account for the semantic difference between x and y (or between x, y and x, x).

III. THE DENIAL OF SEMANTIC ROLE

In stating the antinomy of the variable, we have presupposed that variables *have* a semantic role; and it might be thought that this is the root cause of our difficulties. For it might be thought that our understanding of variables is inseparable from their role in quantification or other forms of variable-binding and that any attempt to explain the role of *free* variables, apart from their connection with the apparatus of binding, must therefore fail.

There is a familiar approach to the semantics of predicate logic that appears to lend some support to this point of view. For in attempting to provide a semantics for quantified sentences, we face a problem that is in some ways analogous to our antinomy. We wish to assign a

semantic value to a quantified sentence, such as '$\exists x(x > 0)$'; and we naturally do this on the basis of the semantic value assigned to the open sentence '$x > 0$' that is governed by the quantifier. But, given that '$x > 0$' and '$y > 0$' are mere notational variants, they should be assigned the same semantic value; and so we should assign the same semantic value to '$\exists x(x > 0)$' and '$\exists x(y > 0)$'—which is clearly unacceptable.

Now one solution to this problem is to deny that the semantic value of '$\exists x(x > 0)$' is to be assigned on the basis of the semantic value assigned to '$x > 0$'; and once this line is adopted, then consistency demands that we never appeal to the semantic value of an open expression in determining the semantic value of a closed expression. In other words, the semantics for closed expressions should be "autonomous" in the sense of never making a detour through the semantics of open expressions.

There are two main ways in which autonomy of this sort might be achieved. I call them the *instantial* and the *algebraic* approaches respectively. According to the first, the semantic value of a quantified sentence such as '$\exists x(x > 0)$' is made to depend upon the semantic value of a closed instance '$c > 0$' (for example, '$3 > 0$'). The intuitive idea behind this proposal is that, given an understanding of a closed instance '$c > 0$', we thereby understand what it is for an arbitrary individual to satisfy the condition of being greater than 0 and are thereby in a position to understand what it is for some individual or other to satisfy this condition.[4]

According to the second approach, the semantic value of a quantified sentence such as '$\exists x(x > 0)$' is made, in the first place, to depend upon the semantic value of the corresponding λ-term '$\lambda x(x > 0)$', denoting the property of being greater than 0. Of course, this merely pushes the problem back a step, since we now need to account for the semantic value of the λ-terms. But this may be done by successively reducing the complexity of the λ-terms. The semantic value of $\lambda x{\sim}(x > 0)$, for example, may be taken to be the "negation" of the semantic value of $\lambda x(x > 0)$, while the semantic value of $\lambda xy(x \leq y \vee y \leq x)$ may be taken to be the "disjunction" of the semantic values of $\lambda xy(x \leq y)$ and $\lambda xy(y \leq x)$. In this way, the λ-bindings may be driven inwards to the atomic formulas of the symbolism and their application to the atomic formulas may

[4] Cf. Michael Dummett, *Frege: Philosophy of Language* (Cambridge: Harvard, 1973), pp. 15–16.

then be replaced by the application of various "algebraic" operations to the properties or relations signified by the primitive predicates.[5]

In discussing these proposals, it is important to distinguish between two rather different questions. The first is whether they are plausible or even viable. Can a semantics of the proposed sort be given and, if it can, then is it faithful to the way we actually understand the given symbolism? The second question is whether open expressions should be taken to have a semantic role (which it might then be part of the aim of semantics to capture).

Of course, if there is an autonomous semantics for closed expressions, then that deprives us of one reason for thinking that open expressions have a semantic role, since they are not required to have a semantic role in order to account for the semantics of closed expressions; and I suspect that many philosophers who have been attracted to the idea of an autonomous semantics for closed expressions have been inclined, on this basis, to reject a semantic role for open expressions. It seems to me, however, that there are strong independent reasons for thinking that open expressions do indeed have a semantic role.

The intuitive evidence for this appears to be overwhelming. Surely, we are inclined to think, it is at least part of the semantic role of an open term to represent a range of values. It is part of the semantic role of the term '$2n$', for example, to represent any even number and part of the semantic role of the term '$a + b$' to represent any complex number. Just as it is characteristic of a closed term such as '2.3' to represent a particular individual, so it is characteristic of an open term, such as '$2n$', to represent a range of different individuals; and just as the representation of a particular individual is a semantic relationship, so is the representation of a range of individuals. We would therefore appear to have as much reason to regard the representation of a range of individuals as a part of the semantic role of an open term as we have to regard the representation of a particular individual as part of the semantic role of a closed term.

But the opponent of semantic roles for open expressions is unlikely to be impressed with this evidence. For he may argue that, in so far as we think of an open expression as having a semantic role, it is because we think of its variables as being implicitly bound. There is perhaps no need to think of them as being bound by a *quantifier*,

[5] George Bealer is the leading contemporary advocate of this approach—see, for example, his "Completeness in the Theory of Properties, Relations, and Propositions," *Journal of Symbolic Logic*, XLVIII, 2 (1983): 415–26.

thus there is no need to think of us as always using these expressions to express thoughts, but there must at least be some variable-binding operator in the background by means of which the supposed semantic role of the open expression is to be understood. So, for example, the term '$2n$' might be understood, in so far as it is seen to have a semantic role, as doing duty for the set-term '{$2n$: n a natural number}' or for the λ-term '$\lambda n.2n$' (denoting the function from each number to its double).

I do not regard this account of the alleged semantic role of open expressions as at all plausible. The alleged semantic role of open expressions, in terms of its representing a range of values, would appear to be perfectly intelligible quite apart from the possible connection with variable-binding. Indeed, some philosophers have supposed that certain types of variable—or variable-like expression—might not even be subject to quantification or other forms of variable-binding. The schematic letters '*A*' and '*B*' of W.V. Quine,[6] for example, are meant to "stand in" for the sentences of some language and yet are not bindable on pain of supposing that the sentences are names for some special kind of entity. These schematic letters, as much as regular variables, are subject to the antinomy and yet would appear to have an independent semantic role. Of course, Quine might be mistaken in his reasons for thinking that schematic letters are not bindable, but he is surely not mistaken in thinking that their semantic role can be understood apart from the connection with quantification or other forms of variable binding.

The opponent of semantic roles for open expressions also faces the awkward issue of saying what the implicit binding should be taken to be. One wants to say that a term '$2n$' indifferently represents all even natural numbers. Our opponent says that the term can only be regarded as having a semantic role in so far as it is implicitly bound. But by what? Two obvious candidates, when no sentential context is at hand, are given by the set-term '{$2n$: n a natural number}' and the λ-term '$\lambda n.2n$'; a less obvious candidate is the *arbitrary even number*.[7] Thus it must be supposed that an implicit reference and an implicit ontological commitment is made to the set or function or arbitrary object, or something else of this sort. But this is both arbitrary and gratuitous. For there is no reason to suppose that the implicit refer-

[6] As in §1.5 of *Methods of Logic* (Cambridge: Harvard, 1982).

[7] As in my article, "A Defence of Arbitrary Objects," *Aristotelian Society*, Supplementary Volume LVII (1983): 55–77 (reprinted in Fred Landman and Frank Veldman, eds., *Varieties of Formal Semantics* (Cinnaminson: Foris, 1984), pp 123–42), and in my book, *Reasoning with Arbitrary Objects*.

ence is to one of these entities as opposed to the other and it appears entirely irrelevant to our use of the term that it should carry any such implicit reference or commitment.

A perhaps even more decisive objection to the position arises from the consideration of semantic relationships. Not only do open expressions appear to have semantic roles, they also appear to enter into semantic relationships. For example, the "value" of the term '$n + 1$' is always greater than the "value" of 'n', though not of 'm'. But how is our opponent to account for these apparent semantic relationships? If he takes the variables of each term, taken on its own, to be implicitly bound, he is sunk: for then '$n + 1$' will signify the successor function, say, and 'n' the identity function, and so we will lose the special semantic relationship that holds between '$n + 1$' and 'n' as opposed to '$n + 1$' and 'm'. He must therefore take the variables of the two terms to be somehow simultaneously bound. He must say something like: what accounts for the apparent semantic relationship between '$n + 1$' and 'n' is the fact that the quantified sentence '$\forall n(n + 1 > n)$' is true. But it seems bizarre to suppose that one must create this artificial context in which both terms occur in order to explain the semantic relationship between them. What kind of strange semantic relationship between the terms is it that can only be explained by embedding them within a richer language? Indeed, the proposed explanation of the semantic relationship presupposes that the relevant semantic features of the terms are preserved when they are embedded in the context of a single sentence; and so unless we had some independent way of explaining what that semantic relationship was, we would have no way to say what the presupposition was or whether it was correct.

If this is right, then the independent semantic role of open expressions is not to be denied and the antinomy is not to be solved by denying that they have such a role.

IV. THE INSTANTIAL APPROACH

I have so far left open the question of whether there might be an autonomous semantics for closed expressions, one not taking a detour through open expressions. But it seems to me that no such semantics is viable—or, at least, plausible. That this is so is not important for solving the antinomy, but it is important for understanding the significance of its solution. For if a semantics for quantification or for other forms of variable binding must proceed by providing a semantics for open expressions then it should provide, in particular, a semantical account of free variables; and so no such account can be considered satisfactory unless it shows how the antinomy is to be solved. Thus

the failure of existing semantics to solve the antinomy is a mark of their inadequacy; they do not properly, or fully, reveal our understanding of variables.

As I have mentioned, there are two main forms of autonomous semantics, the instantial and the algebraic, though the reasons for thinking them unsatisfactory are somewhat different. Let us begin with the instantial approach, which I regard as a closer approximation to how a satisfactory semantics should proceed.

According to this approach, it will be recalled, a closed quantified sentence, such as $\exists x B(x)$ is to be understood on the basis of one of its instances $B(c)$—the intuitive idea being that from an understanding of $B(c)$, we may acquire an understanding of what it is for an arbitrary individual to satisfy the condition denoted by $B(\ )$ and that, from this, we may then acquire an understanding of what it is for this condition to be satisfied by some individual or other. But although the intuitive idea behind the proposal may be clear, it is far from clear how the proposal is to be made precise.

A certain semantic value is to be assigned to a closed instance $B(c)$ of the existential sentence $\exists x B(x)$. Let us call it a "proposition," though without any commitment to what it is. A certain *condition* is then to be determined on the basis of this proposition. But how? We took it to be the condition denoted by the scheme $B(\ )$ which results from removing all displayed occurrences of the term c from $B(c)$. This suggests that the condition should likewise be taken to be the result of removing all corresponding occurrences of the individual denoted by c from the given proposition; indeed, we are given no other indication of how the condition might be determined. It must therefore be presupposed that there is an operation of "abstraction" which, in application to any proposition and any occurrences of an individual in that proposition, will result in a certain condition or propositional "form" in which the given occurrences of the individual have been removed. Once given such a form, we may then take quantified sentence $\exists x B(x)$ to predicate "existence" of it.

Now a great deal more needs to be said about the operation of abstraction before we have a precisely formulated semantics.[8] But one thing is clear. The use of such an operation in formulating the semantics for predicate logic is not compatible with an extensional approach, one in which we take cognizance only of the individuals

[8] Some hints at how a semantics of this sort might be developed are given in my paper, "The Problem of De Re Modality," in Joseph Almog, John Perry, and Howard Wettstein, eds., *Themes from Kaplan* (New York: Oxford, 1989), pp. 197–272, here pp. 237–38. I hope to deal with the matter more fully elsewhere.

denoted by the closed terms, the extensions of the predicates, and the truth values of the sentences. For if there is to be a meaningful operation of abstraction, then the propositions to which it is to be applied must to some extent share in the structure of the sentences by which they are expressed; they must contain individuals in a way analogous to the way in which the sentences contain terms; and it must make sense to remove the individuals from the propositions in a way that is analogous to the removal of terms from a sentence. But clearly, there is nothing in an extensional approach that would enable us to make sense of such ideas.

So much the worse, one might think, for the extensional approach. But however sympathetic one might be to alternative semantic approaches, it is hard to believe that our current problems lie in the adherence to extensionality. After all, the extensionalist credentials of variables are as good as they get: they simply range over a given domain of individuals without the intervention of different senses for different individuals and without the need for different senses by which the domain might be picked out for different variables. It is therefore hard to see why the addition of variables to a language that was otherwise in conformity with extensional principles should give rise to any essential difficulties. If the extensional project fails, it cannot be because the variables carry some hidden intensional freight.

There is in any case another, more subtle, difficulty with the instantial approach which not even the intensional form of semantics is able to solve. For it is a mistake to suppose that our understanding of the quantified sentence is *derived* from our understanding of a particular instance since there may be no particular instance that we are in a position to understand. Suppose, for example, that the variables range over all points in abstract Euclidean space. Then it is impossible to name any particular point. But if we are capable of understanding any instance of the quantified sentence then, a fortiori, we are incapable of deriving our understanding of the quantified sentence from an understanding of an instance.

Of course, what we really wanted to say was that the understanding of $\exists x B(x)$ should be derived from our understanding of an *arbitrary* instance. We have seen that this should not be taken to mean that our understanding of $\exists x B(x)$ derives from our understanding of some particular instance, though it does not matter which. But then what does it mean?

The only reasonable view seems to be that our understanding of $\exists x B(x)$ should be taken to derive from our general understanding of $B(x)$, that is, from our understanding of the proposition expressed by $B(x)$ for any given value of the variable x. But the idea of a closed

instance then falls to the wayside and we are left with the idea of understanding the quantified sentence $\exists x B(x)$ in terms of the corresponding open sentence $B(x)$. Thus we see that the instantial approach, when properly understood, does not even constitute an autonomous form of semantics.

V. THE ALGEBRAIC APPROACH

Under the alternative autonomous approach, the apparatus of binding is traded in for an algebra of operations, with that apparatus serving, in effect, as a device to indicate how the semantic value for a whole sentence is to be generated from the properties and relations expressed by the primitive predicates that it contains. In this case, there is no difficulty in making the semantics precise or in presenting it in extensional form. But there is a difficulty in seeing how to extend it beyond the standard symbolism of predicate logic.

One difficulty of this sort arises from the use of quantifiers that apply to several variables at once, though in no set order. We might have a quantifier 'always', for example, that implicitly quantifies over all variables "in sight." Let us symbolize it by '$\forall$' (without attached variables) and take '$\forall A$' to indicate that 'A' holds no matter what values are assumed by the free variables occurring in 'A'. The question now arises as to how the algebraist is to understand a sentence such as '$\forall(x + y = y + x)$'.[9] Clearly, he must understand it in terms of the application of a universality operator to a λ-term constructed from '$x + y = y + x$'. But which λ-term? There would appear to be only two options: (i) it is a λ-term, '$\lambda xy\ (x + y = y + x)$' or '$\lambda yx\ (x + y = y + x)$', in which the variables attached to the λ-symbol are taken to occur in a set order; (ii) it is a λ-term, '$\lambda(x + y = y + x)$', in which the λ-symbol is taken, like the quantifier-symbol $\forall$, to apply to all free variables in sight, though in no set order.

But the first option gives a more specific meaning to the sentence than it actually appears to possess. We must therefore either arbitrarily adopt one interpretation over another or attribute to the sentence an indeterminacy in meaning which it does not seem to have. And one should not think that the relevant order might be given by the standard alphabetic order of the variables or by the order in which they occur in the formula that follows the quantifier, for we might take the variables to be symbols—such as '#' or '*'—which are not

[9] Mathematicians sometimes use the strict equality sign '≡' in this way. '$x + y \equiv y + x$', for example, is used to indicate the universal truth of '$x + y = y + x$'. A similar difficulty arises for branching quantifiers, since the different "branches" of the quantifier do not occur in a set order

alphabetized in any given order and the formula might be written in a nonlinear notation from which no set order of occurrence can be discerned.[10]

The second option is clearly more faithful to our semantic intentions. But it plays havoc with the idea of distributing the binders across the logical connectives. For, on pain of reintroducing an arbitrary order on the variables, a λ-term such as $\lambda(x > y)$ will have to symbolize a relation that is neutral between the "biased" relations symbolized by $\lambda xy\ (x > y)$ and $\lambda yx(x > y)$.[11] But this means that, when we push the binder λ through $\lambda(x > y \vee y > x)$ in order to obtain the disjunction of the relations symbolized by $\lambda(x > y)$ and $\lambda(y > x)$, we will lose track of the alignment between the variables in the two disjuncts and will therefore be unable to distinguish, in the way we should, between $\lambda(x > y \vee y > x)$ and $\lambda(x > y \vee x > y)$.

Another difficulty of the same sort arises from the use of modal and other intensional operators. Suppose that we add an operator '□' for necessity to the symbolism for first-order logic and that we take the quantifiers to be "actualist"—ranging, in each possible world, over the objects that exist in that world. Consider now the algebraist's treatment of $\Box(x = x)$ and of $\Box\exists y(y = x)$. In conformity with the algebraic approach, $\lambda x\,\Box(x = x)$ should be understood to signify the result of applying some operation, call it *necessitation*, to the property (of self-identity) signified by $\lambda x(x = x)$ and $\lambda x\Box\exists y(y = x)$ should be understood to signify the result of applying this same operation to the property (of existence) signified by $\lambda x\exists y(y = x)$. But under an actualist interpretation of the quantifier (and hence also of λ-binding), $\lambda x(x = x)$ and $\lambda x\exists y(y = x)$ will be modally indistinguishable—they will be true, in each possible world, of the individuals that exist in that world. And this presumably means that $\Box(x = x)$ and of $\Box\exists y(y = x)$ will be modally indistinguishable—which, of course, they are not, since the first is satisfied by any individual whatever while the second is only satisfied by the individuals that necessarily exist.

The general point is that the success of the algebraic approach depends upon presupposing the truth of certain "distribution" principles and there is no reason, in general, to suppose that such principles will be true. But a satisfactory account of quantification should be

[10] As with the notation for fractions or determinants: $\frac{x}{y} = \frac{x}{y}$ and $\begin{vmatrix} xy \\ xy \end{vmatrix} = \begin{vmatrix} x & y \\ x & y \end{vmatrix}$.

[11] For more on the idea of a neutral relation, see my article "Neutral Relations," *Philosophical Review*, CIX, 1 (2000): 1–33

general and should not depend, for its success, on special features of the language to which it is applied.

Another serious difficulty with the approach is that it requires us, at almost every turn, to make arbitrary decisions about the interpretation of the symbolism, decisions which have no counterpart in our actual understanding of the symbolism. Let me merely give one illustration. When we push the λ-operators through logically complex formulas, we will eventually reach λ-terms of the form $\lambda x_1 x_2 \ldots x_n F y_1 y_2 \ldots y_n$ in which the binder $\lambda x_1 x_2 \ldots x_n$ governs an atomic formula $F y_1 y_2 \ldots y_n$ (and where some of the variables $y_1, y_2, \ldots, y_n$ may, of course, be the same as one another and the same as some of the variables $x_1, x_2, \ldots, x_n$). How then are these terms to be interpreted?

Presumably, in keeping with the algebraic approach, we should take $\lambda x_1 x_2 \ldots x_n F y_1 y_2 \ldots y_n$ to signify the result of applying some operation to the relation $\mathscr{F}$ signified by the predicate F. So, for example, $\lambda x F x x$ will indicate the reflexive version of $\mathscr{F}$, while $\lambda y x F x y$ will signify the converse of $\mathscr{F}$ (at least, if $\lambda x y F x y$ signifies $\mathscr{F}$ itself). But what of the operation itself? It must presumably be determined on the basis of the relative disposition of the variables $x_1 x_2 \ldots x_n$ in the binder and of the variables $y_1 y_2 \ldots y_n$ in the atomic formula. But there are different ways in which this might be done; and nothing to choose between them. Consider $\lambda z x F x z x$, for example. We could take this to be the result of first forming the "converse" $\lambda z x y F x y z$ of $\lambda x y z F x y z$ and then forming the reflexive version $\lambda z x F x z x$ of the converse, or we could take it to be the result of first forming the reflexive version $\lambda x z F x z x$ of $\lambda x y z F x y z$ and then forming the converse $\lambda z x F x z x$ of the reflexive version; and similarly, and to a much greater degree, for other cases. These choices do not correspond to anything in our actual understanding of the symbolism; and so, again, we face the awkward choice of making the interpretation either arbitrarily specific or unacceptably indeterminate.

The algebraic approach is best viewed as an attempt to see the symbolism of first-order logic as something which it is not. What it provides, in effect, is a translation from a language with variables to one without variables. A λ-term such as $\lambda z x F x z x$, for example, may be taken to be equivalent in meaning to a term $\mathrm{Refl}(\mathrm{Conv}(F))$, indicating the application of reflexive and converse operations to the relation signified by F, and all appeal to variables in the target language is thereby made to disappear. But in making the transition to a variable-free notation, not only are we forced to make arbitrary decisions about how the translation should go; we thereby lose what is most distinctive about the use of variables. For instead of being treated as devices of reference, albeit of a special sort, they are treated as more

or less oblique ways to indicate the application of various operations within a calculus of relations. The problem of understanding our use of variables is not explained, but side-stepped.

VI. THE RELATIONAL APPROACH

I now wish to indicate how I think the antinomy is to be solved and how a more satisfactory semantics for the symbolism of first-order logic might thereby be developed.

I agree with the autonomous approach in thinking that the formulation of the antinomy embodies a false presupposition. But the false presupposition lies not in the attribution of a semantic role to free variables but in the presumption that there is conflict between the sameness in semantic role of x and y, on the one hand, and the difference in semantic role of x, y and x, x, on the other ((SS)′ and (SD)′ above).

There are, I believe, two things that stand in the way of our seeing how these attributions of sameness and difference might be reconciled. The first concerns a possible ambiguity in the notion of semantic role. We have already had occasion to distinguish between sameness or difference of semantic role *across* contexts and *within* a given context. But there is, I believe, another ambiguity in the notion of semantic role that might stand in the way of seeing how reconciliation is to be achieved.

This may be brought to light by considering the following argument against reconciliation. Suppose, in conformity with (SD)′, that the semantic roles of the pair x, y and of the pair x, x are not the same. Then it may be argued that the semantic roles of the individual variables x and y cannot be the same, in contradiction to (SS)′. For x, when paired with x, has the same semantic role as x, x whereas y, when paired with x, does not have the same semantic role as x, x. The variables therefore differ in respect of whether their pairing with x gives something with the same semantic role as x, x; and this difference, presumably, is a semantic difference.

It is not to be denied that there is a semantic difference of the presumed sort between x and y. But it is not in this sense that we wish to deny that there is a semantic difference between x and y. To see how this is so, let us distinguish between the *intrinsic* (or nonrelational) and the *extrinsic* (or relational) semantic features of an expression. The intrinsic semantic features of an expression, in contrast to its extrinsic semantic features, do not concern its semantic relationship to other expressions. Thus it will be an intrinsic semantic feature of the predicate 'doctor' that it is true of doctors but not an intrinsic semantic feature that it is synonymous with 'physician'. Likewise, the

intrinsic semantic features of a *pair* of expressions will consist of those semantic relationships between the expressions which do not concern their semantic relationship to yet further expressions. Thus it will be an intrinsic semantic feature of the pair 'doctor', 'physician' that they are synonymous, although not that they are both synonymous with 'licensed medical practitioner'.

Now what the above argument shows is that if there is an intrinsic semantic difference between the pairs of variables x, y and x, x, then there will be an extrinsic semantic difference between the individual variables x and y themselves (one concerning the relationship of each to the variable x). But in asserting that the semantic role of x and y is the same, we only wish to assert that their intrinsic semantic features are the same; and in asserting that the semantic roles of x, y and x, x are different, we mean to assert that their intrinsic semantic features are different. Thus the present difficulty will not arise as long as we always take semantic role to be intrinsic.

The other impediment to reconciliation rests upon a mistake in doctrine rather than upon a failure to recognize a distinction. Let it be granted that the relevant notion of semantic role is both crosscontextual and intrinsic. Still, it might be asked, how can there be a difference in the intrinsic semantic relationships holding between each of two pairs of expressions without there being a difference in the intrinsic semantic features of the expressions themselves? Thus given a difference in relationship between the pairs 'doctor', 'doctor' and 'doctor', 'dentist' (with the one pair being synonymous and the other not), there must be a difference in meaning between 'doctor' and 'dentist'. Similarly, given that there is a difference in the semantic relationships holding between the pairs x, y and x, x, must there not be a difference in meaning between x and y? Indeed, without such a difference, we would have, in each case, a pair of synonymous expressions; and it is then hard to see how they could give rise to a difference in relationship. But if the variables differ in meaning, one must have one meaning and the other another meaning; and this will be a difference in their intrinsic semantic features.

According to this view, then, there can be no difference in intrinsic semantic relationship without a difference in intrinsic semantic feature. All differences in meaning must be attributable to intrinsic differences; and any attempt to reconcile the attributions of semantic sameness and difference is therefore doomed to fail.

It has to be acknowledged that this view of meaning—what we might call *semantical intrinsicalism*—is extremely plausible. But it is false; and a careful examination of the behavior of variables indicates how. Suppose again, to fix our ideas, that we are dealing with a

language that contains the variables x_1, x_2, x_3,.... How then is their semantic behavior to be described? For simplicity, let us suppose that their semantic behavior is to be described entirely in extensional terms, since nothing in our argument will turn upon allowing intensional elements to appear in the description.

Now we should certainly specify the range of values each variable can assume and, of course, as long as the language is "one-sorted," the range of values for each variable will be the same. Now normally, in providing some kind of semantical description of the variables, nothing more is said (perhaps because of the grip of intrinsicalist doctrine). But something more does need to be said. For we should specify not only which values each *single* variable can assume, when taken on its own, but also which values *several* variables can assume, when taken together. We should specify, for example, not only that x_1 can assume the number 2 as a value, say, and x_2 the number 3 but also that x_1 and x_2 can *simultaneously* assume the numbers 2 and 3 as values; and, in general, we should state that the variables take their values independently of one another, that a variable can take any value from its range regardless of which values the other variables might assume.

It is important to appreciate here that it does not follow, simply from the specification of a range of values for each variable, which values the variables can *simultaneously* assume. One might adopt the proposal from Ludwig Wittgenstein's *Tractatus*, for example, and disallow distinct variables from taking the same value; or, at the other extreme, one might insist that distinct variables should always assume the same value (treating them, in effect, as strict notational variants of one another); and there are, of course, numerous other possibilities. Thus the fact that distinct variables assume values in complete independence of one another is an additional piece of information concerning their semantic behavior, one not already implicit in the specification of their range.

Once we have specified the range and the independence in value, we will then have a complete description of the semantic behavior of the variables; there is nothing more (at least at the extensional level) to be said about their role. But if this is so, then it is clear that the intrinsicalist doctrine, *no difference in semantic relationship without a difference in semantic feature*, must fail. For the intrinsic semantic features of any two variables will be the same—it will in effect be given by the specification of their range, whereas the intrinsic semantic features of the pairs x_1, x_2, say, and x_1, x_1 will be different, since the former will assume *any* pair of values from the given range while the latter will only assume *identical* pairs of values. If we are merely informed

of the intrinsic semantic features of two variables, then we cannot tell whether they assume their values independently of one another (should they be distinct) or whether they always assume the same value (should they be the same).

It is thus by giving up the intrinsicalist doctrine, plausible as it initially appears to be, that the antinomy is to be solved. We must allow that any two variables will be semantically the same, even though pairs of identical and of distinct variables are semantically different; and we should, in general, be open to the possibility that the meaning of the expressions of a language is to be given in terms of their semantic relationships to one another.

Formally, the situation is analogous to failures in the identity of indiscernibles. Consider Max Black's example of two distinct but indiscernible spheres.[12] Just as there is no intrinsic spatial difference between the two spheres (once we reject absolute space), so there is no intrinsic semantic difference between two variables. And just as there may be an intrinsic spatial difference between two pairs of spheres—since identical spheres will be coincident whereas distinct spheres will not be, so there may be an intrinsic semantical difference between pairs of variables. From this perspective, then, the lesson to be drawn from the antinomy is that semantics provides another, though less familiar, example of an aspect of "reality" in which things can only be distinguished in terms of their relations to one another, though not in terms of their intrinsic features.

VII. RELATIONAL SEMANTICS FOR FIRST-ORDER LOGIC

I now wish to consider how the relational solution to the antinomy is capable of yielding a more satisfactory semantics for the symbolism of first-order logic. I deal first with the case of free variables, which is somewhat more straightforward, and then consider the complications that arise from the use of bound variables.

I take it to be the aim of a semantics for a given language to account for the semantic behavior of its expressions. This is usually done by assigning a semantic value to each meaningful expression of the language, with the semantic value of a complex expression being determined on the basis of the semantic values of the simpler expressions from which it is syntactically derived. The semantic value of each expression is naturally taken to correspond to its intrinsic semantic role; and so, given the truth of semantic intrinsicalism, the assign-

[12] "The Identity of Indiscernibles," in Michael Loux, ed., *Universals and Particulars* (Garden City, NY: Doubleday, 1970), pp 204–16

ment of semantic values to the expressions of a language should then be sufficient to determine their semantic behavior.

But we have seen that the doctrine of semantic intrinsicalism should be abandoned; there are intrinsic semantic relationships between expressions that are not grounded in their intrinsic semantic role. This means that the aim of semantics should be reconceived. For it is no longer sufficient to assign semantic values to expressions; we should also take account of semantic relationships between expressions that may not be grounded in their intrinsic semantic features. Let us use *semantic connection* as the formal counterpart, within a semantics, to the informal notion of semantic relationship (just as *semantic value* is the formal counterpart, within a semantics, to the informal notion of semantic role). A semantics should concern itself, then, with the assignment of semantic connections, and not merely with the assignment of semantic values.

Indeed, we should go further still. It is not that semantic inquiry, as it is usually conceived, needs to be *supplemented* by an account of semantic connection. For, given that there are irreducible semantic relationships among expressions, one can reasonably expect that the semantic value of certain complex expressions will depend on the semantic connections among the simpler expressions from which it is derived, and not merely on their semantic values. Thus semantic connection should *replace* semantic value as the principal object of semantic inquiry and the aim of semantics should be to determine the semantic connections among expressions on the basis of the semantic connections among simpler expressions.

Following through this strategy, the semantics for a given language will eventually terminate in the "lexical semantics," which accounts for the behavior of those expressions, the "lexical items," that are not derived from any other expressions. However, the lexical semantics, like the semantics as a whole, must now be taken to assign semantic connections to the lexical items, and not merely semantic values. It is on the basis of these "primitive" semantic connections that the semantic connections among all expressions of the language will ultimately be determined. This is the general idea of a relational semantics.

If we are to apply this general idea to the language of first-order logic, then we must first have some conception of what we want the semantic connections to be. This may be obtained by generalizing the notion of a value range for a variable. The value range of a variable is the set of values it is capable of assuming. Similarly, given a sequence of expressions, we may take its value range—or semantic connection—to be the set of sequences of values that the expressions are

simultaneously capable of assuming. So, for example, the semantic connection on the sequence of expressions '$x + y$', '$x > y$', 'z' will include the sequence of values 5, FALSE, 6 (obtained under the assignment of 2 to 'x', 3 to 'y' and 6 to 'z') and the sequence 6, TRUE, 2 (obtained under the assignment of 4 to 'x', 2 to 'y' and 2 to 'z'). It should be noted that the semantic connections are entirely nontypographic; they contain no trace of language and there is therefore no danger of the semantics being implicitly typographic.

We must now show how to determine the semantic connection on any given sequence of expressions—starting with the lexical semantics, for the very simplest expressions, and then successively working through more and more complicated forms of expression. The lexical semantics is, for the most part, straightforward: extensions should be assigned to predicates, denotations to constants, and functions to function symbols. However, we now include variables within the lexicon and so the lexical semantics should also specify the semantic connection on any sequence of variables. Suppose that we are given the sequence of variables x, y, x, y, for example. Then in conformity with our understanding that distinct variables take values independently of one another and that identical variables take the same value, the semantic connection on this sequence should be the set of all quadruples a, b, c, d of individuals from the domain for which $a = c$ and $b = d$. And, in general, the semantic connection on the variables x_1, x_2,..., x_n should be taken to be the set of all n-tples a_1, a_2,..., a_n of individuals from the domain for which $a_i = a_j$ whenever $x_i = x_j$ $(1 \leq i < j \leq n)$. It is at this point that relationism enters the semantic scene.[13]

The semantics must now be extended to complex terms. Let us consider, by way of example, the complex terms $x.x$ and $x.y$. The first should have as its value range the set of all nonnegative reals; and the second should have as its value range the set of all reals whatever. How do we secure this result? If we let the value range of $x.x$ simply be a function of the value range of x and of x, and similarly for $x.y$, then we cannot distinguish between them, since the value ranges of x and of y are the same. However, we take the value range of $x.x$ to be a function of the semantic connection on x, x and the value range of $x.y$ to be a function of the semantic connection on x, y. But these semantic connections differ, as we have seen, the first comprising all identical pairs of reals and the second comprising all pairs of reals whatever. And there is therefore a corresponding difference in the

[13] The lexical semantics should also let us combine the connection on two sequences of expressions—such x, y and $+$, 0—which are independent of one another.

value ranges of $x.x$ and $x.y$, for each will comprise the corresponding set of products and will thereby give us the result we want.

More generally, let us suppose that we have a complex term $ft_1t_2...t_n$ and that we wish to determine the semantic connection on a sequence of expressions involving the term. The general form of such a sequence will be $D_1, D_2,..., D_p, ft_1t_2...,\ t_n, E_1, E_2,..., E_q$ (with $p, q \geq 0$). Now, under the standard approach to semantics, the semantic value of the complex term $ft_1t_2...t_n$ will be determined on the basis of the semantic value f of its function symbol f and the semantic values $a_1, a_2,..., a_n$ (under a given assignment) of its argument-terms $t_1, t_2,..., t_n$; and this semantic value will be the result $f(a_1, a_2,..., a_n)$ of applying the function f to the arguments $a_1, a_2,..., a_n$. This suggests that the semantic connection on the sequence $D_1, D_2,..., D_p, ft_1t_2...t_n, E_1, E_2,..., E_q$ should be determined on the basis of the semantic connection on $D_1, D_2,..., D_p, f, t_1, t_2,..., t_n, E_1, E_2,..., E_q$ (with $ft_1t_2...t_n$ giving way, as before, to $f, t_1, t_2,..., t_n$); and $d_1, d_2,..., d_p, b, o_1, o_2,..., o_q$ will belong to the semantic connection on $D_1, D_2,..., D_p, ft_1t_2...t_n, E_1, E_2,..., E_q$, that is, $ft_1t_2...t_n$ will be capable of taking the value b when $D_1, D_2,..., D_p, E_1, E_2,..., E_q$ take the values $d_1, d_2,..., d_p, o_1, o_2,..., o_q$ just in case, for some individuals $a_1, a_2,..., a_n$ for which $b = f(a_1, a_2,..., a_n)$, $t_1, t_2,..., t_n$ are capable of taking the values $a_1, a_2,..., a_n$ when $D_1, D_2,..., D_p, E_1, E_2,..., E_q$ take the values $d_1, d_2,..., d_p, o_1, o_2,..., o_q$, that is, just in case, for some such individuals $a_1, a_2,..., a_n$, the n-tple $d_1, d_2,..., d_p, f, a_1, a_2,..., a_n, o_1, o_2,..., o_q$ belongs to the semantic connection on $D_1, D_2,..., D_p, f, t_1, t_2,..., t_n, E_1, E_2,..., E_q$.

The above rule is easily extended to the case of atomic formulas; and a similar rule may be given in the case of truth-functionally complex formulas. The semantic connection on $\sim A, E$, for example, will consist of all those pairs π', o for which π, o is a member of the semantic connection on A, E and π' is the "opposite" truth value to π.

Quantifiers raise some additional problems. The obvious way of evaluating a sequence containing a quantified formula—such as $\exists xA(x), E$—is in terms of the connection on $x, A(x), E$. If, for a given semantic value of E, $A(x)$ is true for the assignment of some individual from the domain to 'x', then $\exists xA(x)$ should be taken to be true and otherwise should be taken to be false. Thus the pair TRUE, o will belong to the connection on $\exists xA(x), E$ just in case, for some individual o from the domain, the triple o, TRUE, o belongs to the connection on $x, A(x), E$ (and similarly for the pair FALSE, o).[14]

[14] I have assumed that the quantified formula $\exists xA(x)$ contains no free variables. If it does, then we can only properly evaluate the formula by considering its connection to the free variables that it contains; the above rule is in effect relativized to an assignment of individuals to those variables.

But consider how such a rule applies to a sequence of the form $\exists x A(x)$, x—say to $\exists x(x > 0)$, x (where the domain of quantification is the set of all natural numbers). The pair TRUE, 0 will belong to the semantic connection on $\exists x(x > 0)$, x just in case, for some natural number n, the triple n, TRUE, 0 belongs to the semantic connection on x, $x > 0$, x. But since the first and third variables in x, $x > 0$, x are the same, the first and third components of any triple in its semantic connection will be the same. It follows that no triple of the form n, TRUE, 0 can belong to the semantic connection on x, $x > 0$, x and so the pair TRUE, 0 will not belong to the semantic connection on $\exists x(x > 0)$, x—contrary to our intentions.

The problem is that we do not want the bound occurrences of the variable x in $\exists x A(x)$, x to be "coordinated" with the free occurrence. However, our method of evaluation requires that they be coordinated since $\exists x A(x)$, x is evaluated in terms of x, $A(x)$, x. What makes the problem especially acute is that we wish to subject $\exists x(x > 0)$, x to essentially the same method of evaluation as $\exists y(y > 0)$, x, since they are mere notational variants of one another, and yet we also want to subject $\exists y(y > 0)$, x to the straightforward evaluation in terms of y, $y > 0$, x.

The way out of the difficulty, I believe, is to give up the assumption that all occurrences of the same variable should be treated in the same way. We have so far assumed that different free occurrences of the same variable should always take the same individual as value; and this is, indeed, a reasonable default assumption to make. However, when a free occurrence of a variable was originally bound, only becoming free in the process of evaluation, then we should no longer assume that it is coordinated with those occurrences of the same variable that were either originally free or had their origin in a different quantifier.

This means that, for the purpose of evaluating a sequence of expressions, we should explicitly indicate which of the free occurrences of a given variable are to be coordinated and which not. In the sequence x, $x > 0$, x, for example, we should distinguish between the cases in which none of the occurrences of x are to be coordinated, all are, or just two are. (The reader might picture these coordinating links as lines connecting one occurrence of the variable to another.)

This then provides us with the means of evaluating $\exists x(x > 0)$, x in terms of x, $x > 0$, x, in strict analogy with the evaluation of $\exists y(y > 0)$, x in terms of y, $y > 0$, x. However, the variables in the evaluating sequences should be subject to the appropriate pattern of coordination; the first two occurrences of x in x, $x > 0$, x should be coordinated with one another (though not with the third occurrence) and the

first two occurrences of y in y, $y > 0$, x should likewise be coordinated. In general, whenever we are evaluating a quantified formula $\exists x A(x)$ in terms of its components x, $A(x)$ within the context of a sequence, the designated occurrences of x should be taken to be coordinated with one another, though not with any other occurrences of x that might happen to be present in the sequence.

Modest as this proposal might appear to be, its development calls for fundamental revisions in our previous formulation of the semantics. In the first place, the syntactic object of evaluation will no longer be a sequence of expressions but a *coordinated* sequence of expressions. This is a sequence of expressions $E_1,\ldots, E_n$ along with a *coordination scheme* $\mathfrak{C}$ which tells us when two occurrences of the same variable are to be coordinated (formally, a coordination scheme is an equivalence relation on the occurrences of variables in the sequence which only relates occurrences of the same variable). In the second place, the lexical rule for variables must be modified. Instead of requiring that all occurrences of the same variable should receive the same value, we should only require that they receive the same value when they are coordinated. With these two changes in place, the semantics may then be stated much as before.

The relational semantics has several clear advantages over its rivals. First and foremost, it embodies a solution to the antinomy: the intrinsic semantic features of x and y (as given by the degenerate semantic connections on those variables) are the same, though the intrinsic semantic features of the pairs x, y and x, x (again, as given by the semantic connections on those pairs) are different. The semantics is also more satisfactory, in various ways, *as* a semantics. In contrast to the autonomous approaches, it assigns a semantic role to open expressions; in contrast to the instantial approach, it can be given an extensional formulation; and in contrast to the algebraic approach, it is based upon a credible direct method of evaluation. And finally, in contrast to the Tarski semantics, it is nontypographic; by going relational, we avoid having to incorporate the variables themselves into the very identity of the entities that the semantics assigns to the open expressions of the language.

The reader may be worried by the complexity of our account, especially in comparison with the Tarski semantics. But one should bear in mind that the more abstract a semantics for a given language, the more removed it is from the underlying syntax, the more complicated it is likely to be. Indeed, it seems to me that the complexity of our own account is of a desirable sort. For it appears not to be a mere artifact of the approach but to reflect a genuine complexity in our

understanding. The relational approach forces us to take seriously the pattern of coordination among the variables; these patterns must be explicitly used as syntactic inputs to the semantic method of evaluation. And it is hard not to believe that their semantic significance, as revealed by the semantics, is not somehow integral to our understanding of the variables. From among the alternatives, it is only the algebraic approach that takes explicit account of the phenomenon of coordination; but it does this in a manner that is completely artificial and not in keeping with the variable's referential role.

VIII. IMPLICATIONS

Our approach to the semantics of variables has various broader implications of which I should provide a brief discussion.[15]

It is important, in this connection, to distinguish between the present version of relationism and the more familiar doctrine of semantic holism. Our own version of relationism is meant to serve as an *extension* to the usual forms of representational semantics. We take it to be a legitimate task of semantics to assign referents or senses to expressions and then consider whether semantics, as so conceived, might also require the postulation of semantic relationships. Semantic holism, on the other hand, is meant to serve as an *alternative* to the usual forms of representational semantics. The task of assigning referents or senses to expressions is abandoned in favor of the task of providing an account of their inferential or "conceptual" role. Thus the two approaches are relational in very different ways, with the one staying squarely within the representational framework and with the other locating the relational ideas outside of that framework.

As I have already indicated, the proper semantic treatment of variables provides a clear and convincing refutation of the doctrine of semantic intrinsicalism. But semantic relationism might not be so interesting, even as a general thesis, if the case of variables provided the only instance of its truth. There are, however, many other ways in which the doctrine can be applied.

One of the most interesting, to my mind, concerns the distinction between what one might call "strict" and "accidental" coreference. A Millian will claim that two coreferential proper names, such as 'Cicero' and 'Tully', will have the same intrinsic semantic features, since there is no more to their meaning than their referents. However, it is still possible for a Millian to distinguish the semantic relationship holding

[15] I hope to discuss these more extensively in a forthcoming book *Reference, Relation, and Meaning* (to be published by Basil Blackwell).

between 'Cicero' and 'Cicero' (which is one of *strict* coreference) from the semantic relationship holding between 'Cicero' and 'Tully' (which is one of *accidental* coreference). Thus strict or accidental coreference will be semantic relationships for the Millian that are not grounded in the intrinsic semantic features of the expressions between which they hold. Moreover, once this distinction is made, the Millian has new means of dealing with Frege's puzzle, with Saul Kripke's puzzle concerning belief,[16] and with various related puzzles concerning the relationship between semantics and cognition. Indeed, by marrying relationism with Millianism, we may obtain a much more plausible version of direct reference theory, one that is not subject to the usual Fregean objections. In effect, relation serves as a substitute for sense.

A related distinction may be drawn between *strict* and *accidental* synonymy. Thus 'doctor' and 'doctor' will be strictly synonymous, while 'doctor' and 'physician' will only be accidentally synonymous. This distinction should be accepted by the Fregean, I believe, and not merely the Millian; and it provides us with a new, and more plausible, way of dealing with Benson Mates's puzzle[17] and with G. E. Moore's paradox of analysis. These puzzles can be solved without appeal either to a difference in higher-order sense or to a difference in the expressions themselves.

Relational ideas are also important to the project of constructing a detailed semantics for ordinary language. The most obvious application is to anaphora, which should be treated as a form of strict coreference (or synonymy). But there are numerous correlative terms and constructions which also appear to call for a relational treatment. Coordinated uses of 'Mr. Smith' and 'Mrs. Smith', for example, should be taken to refer to husband and wife; and in a construction such as 'the more the merrier', it is clear that the semantic value of 'the more' should somehow be coordinated with the semantic value of 'the merrier'.

Finally, there are further applications to be made to logic itself. For the apparatus of coordination provides us with much greater flexibility in formulating different forms of semantics for the use of variables. It can accommodate a form of "dynamic" semantics, for example, one in which it is possible to "quantify out"; and it enables us to provide a much more plausible account of referential quantification

[16] "A Puzzle about Belief," in Avishai Margalit, ed., *Meaning and Use* (Boston: Reidel, 1979), pp. 239–83.

[17] "Synonymity" in Leonard Linsky, ed., *Semantics and the Philosophy of Language* (Urbana Illinois UP, 1972), pp. 111–38.

into intensional contexts than the standard approaches, since coordination between variables can now be made to "track" significant semantic distinctions. Thus we see that, quite apart from its general relevance to the philosophy of language, semantic relationism has detailed implications for the development of semantics in a wide variety of areas.

KIT FINE

New York University

BOOK REVIEWS

Freedom Evolves. DANIEL C. DENNETT. New York: Viking, 2003. xiii + 347 p. Cloth $24.95.*

Freedom Evolves is a vigorous and lively defense of a naturalistic, compatibilistic notion of freedom (and moral responsibility). It shows how these sorts of phenomena can be given evolutionary explanations; also, it exhibits the evolution of Daniel Dennett's own account of freedom, which he presented in his important contribution to the free will debates, *Elbow Room: The Varieties of Free Will Worth Wanting*.[1] It responds to various objections to his sort of compatibilist account, comments on some contemporary philosophical literature, and also exhibits the connections between Dennett's approach to freedom and his influential accounts of consciousness, the self, and Darwinian evolutionary theory. Notably, Dennett connects the philosophical issues pertaining to free will with the insights and methodologies of contemporary psychologists such as Daniel Wegner and George Ainslie, economists such as Robert Frank, and biologists such as Richard Dawkins, Jared Diamond, Edward O. Wilson, and David Sloan Wilson. This is a synoptic and compendious book, not likely to be appealing to the more meticulous analytic philosophers. Its target, I would think, is primarily a more general audience of intellectuals and scholars in other fields, although even the most fastidious philosopher may find some intriguing, suggestive, and provocative observations.

The package of theses defended by Dennett is fairly familiar. Dennett argues that a "real" and robust sort of freedom is entirely compatible with a thoroughgoing, naturalistic causal determinism. The following statement nicely (if somewhat vaguely) captures Dennett's presuppositions and methodology: "My fundamental perspective is *naturalism*, the idea that philosophical investigations are not superior to, or prior to, investigations in the natural sciences, but in partnership with those truth-seeking enterprises, and that the proper job for philosophers here is to clarify and unite the often warring perspectives into a single vision of the universe" (15). He contends that nothing

* I am grateful for helpful comments by Al Mele and Manuel Vargas
[1] Cambridge: MIT, 1984.

0022-362X/03/0012/632–37 © 2003 The Journal of Philosophy, Inc.

in our practices or in critical reflection on them entails that we need to assume that indeterminism is true, or that we have (what he takes to be) mysterious, irreducible powers to "agent-cause" our behavior, or that materialism is false, and so forth. Further, the freedom we care about and our practices of holding each other morally responsible do not presuppose a kind of possibility to do otherwise that holds fixed all the details of the past and the natural laws. Nothing of significance is lost, then, if we give up this "libertarian" notion of freedom.

Although some theorists have evidently concluded from Dennett's previous defenses of materialistic accounts of mental states, consciousness, and the self, that "there is no room for Free Will" (on such an account), Dennett points out that this is a mistake. He develops a series of thought experiments, or, in his much-preferred terminology, "Just-So Stories," that employ insights and analytical approaches from evolutionary biology, economics, and psychology to show how one can defend a strong notion of freedom even in a materialistic, causally deterministic world. Of course, the tools for defending such a view will include the recognition of the extraordinary *complexity* of (physical) creatures like us, and the appropriateness of *different levels of description of*, or *different stances to*, such organisms.

Dennett tells an elaborate Just-So Story in which initially simple organisms evolve into complex creatures that have and respond to reasons, make choices, and have freedom, even in a causally deterministic world. A landmark event in this Story takes place when the creatures develop the ability to speculate about the future, to consider possible threats that jeopardize their interests and plans, and to make choices that help them to *avoid* those projected future threats. "Evitability," then, is a significant achievement, and clearly confers an important evolutionary advantage. And, Dennett argues, the sort of evitability in question is not understood in terms of the "possibility to do otherwise, holding all the details of the past and natural laws fixed"; it is perfectly compatible with causal determinism.

Much of what is of value in this book is contained in the details of the Just-So Stories. There are also intriguing critical discussions of the indeterministic theory of free will presented by Robert Kane, and some of the views about the human will suggested by the neuroscientist, Benjamin Libet. Dennett builds on previous work that suggests a helpful kind of skepticism about simple thought experiments in general, and, in particular, those that posit the weird inculcation of beliefs, desires, and other mental states (by nefarious neurosurgeons and their ilk). Dennett here explores Alfred Mele's example of Ann and Beth, who are allegedly psychologically isomorphic except that Beth has had crucial mental states induced by a neuroscientist. Al-

though Mele's point was primarily to defend a *historical* account of autonomy, Dennett addresses some complexities in such putative examples of induction of mental states.

I agree with Dennett about a lot: that causal determinism is compatible with a sort of freedom and also with moral responsibility, that moral responsibility is properly understood in terms of a kind of sensitivity to reasons, that this sensitivity can be present even in the absence of libertarian freedom, and so forth. I am also comfortable with a "naturalistic" methodology (on most plausible understandings of "naturalism"), although I suppose that I would want something slightly different: an approach that is *compatible* with naturalism, even if it does not *entail* it.[2]

Nevertheless, I have some significant disagreements with Dennett. One of these disagreements pertains to what might be called "libertarian freedom." Both of us hold that such freedom is not necessary for moral responsibility. But I take seriously the libertarian's wish to have such freedom, and take it as significant that causal determinism evidently rules out such freedom; Dennett, in contrast, dismisses this sort of freedom as uninteresting. For example, Dennett says:

> Many philosophers have assumed without specific argument that when we ask a question about what was possible, we are—and should be—interested in knowing whether, in *exactly* the same circumstances, the same event would recur. We have argued that in spite of its traditional endorsement by philosophers, this policy is *never* followed by serious investigators of possibility and is, in any event, unmotivated: It *couldn't* give you an answer that could satisfy your curiosity. The burden now rests with those who think otherwise to explain why 'real' possibility demands a narrow choice of X [the set of possible worlds relevant to the possibility-claim]—or why we should be interested in such a concept of possibility, regardless of its 'reality' (82).

It is completely uncontroversial that there are various kinds of possibility, and thus various corresponding sets of possible worlds (relevant to the possibility-claims in question). Further, I absolutely agree with Dennett that different contexts of inquiry will render different notions of possibility relevant, and there are important contexts of inquiry to which broader, rather than narrower, notions of possibil-

[2] The term, 'naturalism', is bandied about quite promiscuously, and it is often unclear what a theorist means by this term. I would not want a theory of moral responsibility that entails or presupposes materialism about the mind, for example (although I myself am inclined to accept materialism). It would also be unattractive to make it a *requirement* of an adequate theory of moral responsibility that God does not exist.

ity are appropriate. For example, we are often interested in what is physically possible (not just what is physically possible, given a fixed past and set of natural laws). We are interested in what sorts of things are consistent with the laws of nature, and since we do not know the details of the total true description of the past, we will sensibly inquire, in many contexts, simply about what kinds of things are consistent with the laws of nature. Our epistemic limitations, together with our pragmatic goals, make it perfectly appropriate for us to concern ourselves with broad notions of physical possibility in our scientific theorizing.

But it does *not* follow that the narrower notion of possibility is not important, and, in particular, that it is not important to our freedom. When we think of ourselves as having various pathways genuinely open to us, it is very plausible to suppose that we are picturing various paths that branch off a fixed path, holding the natural laws fixed. That is, when we think of ourselves as being free to pursue a particular path into the future, we explicitly or implicitly think of that path as an extension of the actual past (holding the laws of nature fixed). We typically do not think of ourselves as having the power so to act that the past would have been different from the way it actually was, or so to act that the laws of nature would have been different. So, the picture is this: the various accessible pathways into the future branch off a fixed past (and contain the actual natural laws), and the reasons relevant to our practical reasoning are reasons that obtain only along the accessible pathways. A benefit that exists along some path is pertinent to my practical reasoning only if I can "get there from here."

The above picture lies behind various formalizations of a "Basic Argument" for the incompatibility of causal determinism and free will, in the sense that involves genuine access to alternative possibilities.[3] One can deny the picture—that is, one can deny that the past or the laws are fixed in the relevant sense.[4] So it is not uncontroversial that the picture corresponds to the correct account of possibility, in the free will sense. But it is an extremely plausible, attractive, and natural picture. I find it hard to give up the idea that my freedom, in a given context, consists in my power to add to the given past, holding the natural laws fixed. The picture of "multiple pasts"—

[3] I discuss at greater length Dennett's critique of the Basic Argument in my contribution to the "Author-Meets-Critics" session on the book at the American Philosophical Association Pacific Division Meetings, March 2004 (Pasadena, California).

[4] I discuss such moves in John Martin Fischer, *The Metaphysics of Free Will* (Cambridge: Blackwell, 1994), pp. 67–86.

different pasts for each path I could take—is as unintuitive as it is unlovely. And the Humean idea that I can write the laws of nature (and that there are no laws of nature until the end of time) strikes me as somewhat bizarre. This picture of the structure of free will possibility is precisely what must be abandoned by the compatibilist about causal determinism and free will, in the sense that involves genuine access to alternative possibilities.

Perhaps then it is not surprising that philosophers have been concerned about precisely this sort of freedom for thousands of years. (By the way, the Basic Argument goes back to Origen—not to be confused with one of Dennett's favorite topics, *The Origin of Species*!) I, like Dennett, am a compatibilist about moral responsibility and causal determinism. But I do not think that compatibilism makes progress unless it takes the libertarian's concerns seriously. It does not make progress if it abruptly dismisses a set of concerns that not only resonates deeply with common sense, but with thousands of years of philosophical analysis.

I wish to end with a few reflections on features of Dennett's methodology. As is well known, Dennett does not trust philosophical "thought experiments," about which he expresses considerable skepticism. He is particularly wary of thought experiments involving the artificial induction or inculcation of mental states (by "nefarious neurosurgeons"—or even nice ones, presumably). But he engages (especially in this book) in his own story telling. Why exactly are Dennett's Just-So Stories any different from traditional philosophical thought experiments? In both cases there is significant simplification and abstraction. Dennett gives hegemony to the fictional stories that are "dressed up" with technical or "scientific" features, but why? After all, these stories (like the traditional philosophical thought experiments) are abstractions, the conclusions of which need to be applied with delicacy and sensitivity to the nuances of the real world. In my view, a philosophical thought experiment *does* need to be handled with care and extreme caution; but this does not imply that such thought experiments cannot be useful and illuminating.[5]

Philosophical thought experiments are stories that invite us to consider our principles more carefully. In some instances, they force us to give a more nuanced analysis that is sensitive to the complexity of the relevant terrain. I often feel that Dennett's discussion could bene-

[5] See Fischer and Mark Ravizza, "Introduction," in Fischer and Ravizza, eds., *Ethics: Problems and Principles* (Fort Worth, TX: Harcourt, 1992), pp. 12–16 and 45–51; and Fischer, "Stories," *Midwest Studies in Philosophy*, XX (1996): 1–14.

fit from a subtler, more articulated view of the target phenomena. For just one example, Dennett often speaks of "freedom," without distinguishing what I think are importantly different *kinds* of freedom. If one abruptly dismisses the thought experiments (as well as other, more traditional forms of philosophical analysis), one can fail to see the complexity of the notions of freedom and responsibility.[6]

Dennett is at his best when he is essentially challenging philosophers to be broader in our intellectual gaze, rendering philosophical theorizing continuous with cutting-edge work in the social and natural sciences. But evidently he feels that the need to avoid parochialism is asymmetric. Here was an opportunity to introduce a broader intellectual audience to the extraordinarily lively and rich contemporary philosophical debates about free will and moral responsibility, but the opportunity was lost.

JOHN MARTIN FISCHER

University of California/Riverside

[6] I do not suppose that we should do without the sorts of scientific approaches and models invoked by Dennett. I simply do not think we should abandon philosophical analysis and abruptly discard philosophical thought experiments. The Rawlsian notion of seeking a *wide* reflective equilibrium seems particularly appealing here—see Fischer and Ravizza, *Responsibility and Control* (New York: Cambridge, 1998), pp. 10–11.

God, Locke, and Equality: Christian Foundations in Locke's Political Thought. JEREMY WALDRON. New York: Cambridge University Press, 2003. xii + 263 p. Cloth $60.00.

In 1969, John Dunn[1] had tried to show that Locke's political theory was so saturated with religious assumptions derived from a particular variety of Protestant Christianity that, as a source of arguments that would satisfy our modern secular tastes, it had become largely irrelevant. Jeremy Waldron has now turned Dunn's case on its head. He argues that our modern belief in something like basic equality *requires* a religious foundation, as Locke demonstrated so clearly. Explicitly distancing himself from his earlier positions on the foundation of liberal rights, Waldron here is a critic of secular reasoning whose moral vocabulary he finds inadequate for a full-fledged defense of universal human equality. I would have been happy to call this an idiosyncratic book, but I am not sure it is.

The reading of Locke that Waldron endorses is as follows. An argument for basic equality based on some notion of the essential identity of the human species as distinct from other species is, according to Locke, unacceptable, because distinctions between species are not natural but merely conventional, "made by men." Instead, Locke argues that there are real resemblances between humans as corporeal rational beings, because all have the capacity for abstraction that can lead them to knowledge of God. This capacity for abstract moral knowledge is not a philosophical skill; ordinary people have this moral capacity because they can have faith in revelation. Since all humans have this capacity, they are all God's special creatures and must be treated as equals. This, of course, is a theologically grounded defense of equality. Waldron claims it is a stronger and more reasonable defense than anything secularists can offer today.

Relying on this core argument, Waldron defends virtually the entire corpus of Locke's political theory of equality. He refutes, largely on the strength of Richard Ashcraft's research, the suggestion that Locke's commonwealth was open only to property-owners and that laborers

[1] *The Political Thought of John Locke: An Historical Account of the Argument of the Two Treatises of Government* (New York: Cambridge, 1969).

0022-362X/03/0012/638–41 © 2003 The Journal of Philosophy, Inc.

were meant to be left out (the specter of C.B. Macpherson still haunts Locke studies). He defends Locke's argument that children must be represented by the will of their parents, because although children are born *to* equality, they are not born *in* it. (He does not even consider the use of this argument for keeping entire peoples in a state of dependency; it is doubtful that Locke would have disagreed with, say, J.S. Mill two hundred years after him when the latter claimed that the Irish or Indians or Africans could not represent themselves because they were moral infants.) He agrees that Locke's exceedingly harsh theory of punishment of criminals, his ferocious portrayal of self-defense, or his support for the enslavement of people defeated in a just war, are inconsistent with the idea of basic equality. But Waldron suggests that these arguments in Locke are not carefully thought out. He concedes James Tully's charge that Locke's idea of mixing one's labor with the land as the highest moral claim to property served as a justification for expropriating Native Americans, but disagrees that this was tantamount to regarding them as less than fully human. Locke, says Waldron, took the duty of survival by the fruits of one's labor as a serious, theologically grounded, moral principle and provided an argument why cultivation by white settlers was better for all, including Native Americans. We may not like Locke's argument, but he was not, says Waldron, dehumanizing the native.

The only place where Waldron finds Locke completely unacceptable is on his argument about the natural subjection of wives to their husbands. This is "an embarrassment for his general theory of equality" (40) because, contrary to the allegations of feminist critics such as Carole Pateman, Locke did not, insists Waldron, restrict his domain of equality to men only. "I have no tidy resolution to offer," he says, and leaves it at that.

Throughout the book, Waldron takes numerous digs at the school of Cambridge scholars who insist on reading texts in their historical context. Yet he does not refrain from using such historicist arguments himself in, for instance, distinguishing between an early Locke and a mature Locke, or using contextual material to speculate on what Locke might have meant by a particular expression. Most extraordinary is his indignant outburst against those who suggest that Locke's personal involvement with American slavery is a blot on his moral integrity as a champion of equality. This charge, he says, flows from our own notions of career and integrity, where political action is nothing more than signing an open letter or writing in the *New York Review of Books*. In Locke's times, political actions carried serious personal risks and we cannot judge his moral integrity by our own

comfortable standards. (Is not that a plea to read texts in their context?)

The mature Locke, says Waldron, was prepared to extend civil rights to Catholics, Jews, Muslims, and pagans, unless they bore allegiance to a foreign power. The only people he consistently refused to include in the commonwealth were atheists. Waldron explains this refusal by restating the theological ground of Locke's theory of basic equality: all humans must be treated as equal because all have the moral capacity for knowing the truth of God. This the atheist does not admit. Hence, the atheist cannot be trusted to respect the right of equality and so must be excluded from the domain of toleration, even though (we could point out) the slave owner continues to be a respected citizen. Waldron agrees that, "as a bottom-line political position," the exclusion of atheists is not an option for us today. "But," he adds, "the bottom-line is not everything" (235).

> Suppose—as I am inclined to believe—that a commitment to human equality is most coherent and attractive when it is grounded in theological truth, truths associated particularly with the Christian heritage (236). They are matters *we must get to the bottom of*, whether or not the bottom turns out to be controversial and philosophically sectarian (238).

Why? Because basic equality is under challenge today from sophisticated theories that seek to establish political power on an inegalitarian basis—Waldron vaguely mentions "various Nietzscheans and radical Freudians"—and it is necessary to reassert the "ethical inheritance" of the idea of equality in the history of post-Reformation Christianity, even at the risk of being sectarian. It must be shown that when atheists defend equality, they do so only as disingenuous Christians.

I must confess I am mystified by this invocation of an epochal moment in the philosophical history of humanity. I was born in what Locke would have called a pagan culture that traditionally ascribed human inequality to the purposeful design of the Creator. Anthropologists like Louis Dumont theorized this religious ideology as an order of *homo hierarchicus* that recognized the inescapable facts of human inequality but tried to regulate them into a harmonious social whole. As I grew up, I read of Buddhism and Jainism as egalitarian critiques of the caste-divided Vedic religion, but critiques that were grounded in atheist premises because they refused to acknowledge a Creator. I read about the historical influence of Islam as a religion that recognized the equality of all humans without claiming the source of landed property in labor or granting equal civil rights to people of all religions. And I was persuaded, like millions of non-Christian peoples in the contemporary world, to believe in basic equality by the writings

of Rousseau, Tom Paine (deism and all!), and Karl Marx, and by the declarations of the French Revolution and, in our own times, the United Nations. I had come to believe that there had emerged a philosophical arena where arguments about equality (even if transcendental) could be adjudicated without recourse to theological or sectarian assumptions. Waldron, it seems, would want his co-religionists to return to the provincial cesspool of Western Christianity. I await other Western philosophers to tell me that he is wrong.

PARTHA CHATTERJEE

Centre for Studies in Social Sciences/Calcutta
Columbia University

NOTES AND NEWS

The Editors are deeply saddened to report the death of Donald Herbert Davidson, Willis S. and Marion Slusser Professor Emeritus of Philosophy at the University of California, Berkeley. He passed away on August 31st in Berkeley at the age of 86. Davidson was born in Springfield, Massachusetts on March 6, 1917. As an undergraduate, he studied at Harvard University, first English and comparative literature, then classics and philosophy, working with A.N. Whitehead. Davidson served in the United States Navy from 1942 to 1945. Returning for graduate work at Harvard, Davidson completed his dissertation on Plato's *Philebus* and received his Ph.D. in 1949. Davidson taught first at Queens College from 1947 to 1950, then at Stanford University from 1950 to 1967, and at Princeton University from 1967 to 1969. He moved to Rockefeller University and then to the University of Chicago during the 1970s before going to Berkeley in 1981. He was the recipient of numerous awards and fellowships throughout his career. Davidson's first major published work was *Decision-making: An Experimental Approach*, co-written with Patrick Suppes in 1957. Davidson's great philosophical achievement, however, is contained in a body of essays spanning four decades, beginning with the publication of "Actions, Reasons, and Causes" in 1963, and including such works as "Truth and Meaning," "Mental Events," "Radical Interpretation," and "On the Very Idea of a Conceptual Scheme." These extraordinarily influential essays are collected in the volumes *Essays on Actions and Events* (1980); *Inquiries into Truth and Interpretation* (1984); and *Subjective, Intersubjective, Objective* (2001); two more volumes of collected papers are expected to be published.

A measure of Davidson's remarkable achievements in philosophy is the extent to which philosophers who write on his themes have defined their views, often very different views, in terms of his. Initially, his influence took the form of a number of younger philosophers elaborating on the details of a program of semantics for natural language along the lines he had first proposed. Perhaps the most important results of this period are to be found in the excellent work that has been done by linguists and philosophers on the logical form of action sentences. It may well be that this aspect of his work will in time be remembered above all others. But there were many other aspects of great significance and wide impact. Over the years, as the excitement over this program subsided, the broader philosophical positions he took came more fully into view and they displayed his great and varied originality. Even a mere (and selective) glance at some of his arguments and conclusions reveals an impressively large

list, one would have thought a list without peer in number and range—the Tarskian theory of truth to account for linguistic meaning, the embedding of such a theory in the method of radical interpretation, the ambitiously integrated use of this method to solve not just for meaning and belief but for desire or value as well, the distinctively austere vision of the social nature of meaning and intentionality, the notorious and insistent appeal to the principle of charity, the raising of the stakes in the philosophy of language and mind by demanding that such a philosophy square with a credible anti-skeptical epistemology, the ingenious arguments against relativism and skepticism, the startlingly narrow understanding of the meanings of metaphors, the flamboyant denial that there is any such thing as "language," the anti-naturalist claim to the irreducibility of intentional states even as they were said to be identical with physical states, the novel analysis of weakness of will and other forms of irrationality. All of these were presented in papers, well known for their compactness and a frequent gleam of phrase. If they seemed elusive sometimes, that was part of their charm, in a subject whose productions are not renowned for charm.

Davidson had two intellectual heroes, Alfred Tarski and his mentor, W.V. Quine. Their influence was pervasive in much of his writing, and throughout his career he would invoke or address them. At the end of his life, he had just completed a work on the nature of predication, a proposed solution to the ancient problem of the unity of the proposition that owed much, once again, to Tarski. A student could always get him to pay close attention to what he or she was saying if it could be presented as bearing some relation to either Tarski or Quine. There was something touching about this loyalty in a man whose loyalties were otherwise sturdily focused on the framework of his own ideas and a life-long determination to apply that framework and push those ideas as far as they would reach. Such intellectual determination and power is rare and it will be long admired.

NEW BOOKS

ADAMSON, GREGORY DALE: *Philosophy in the Age of Science and Capital.* New York: Continuum, 2002. 169 p. Cloth $105.00, paper $24.95.

ALANEN, LILLI: *Descartes's Concept of Mind.* Cambridge: Harvard, 2003. xv, 355 p. Cloth $65.00.

AUSTIN, JAMES H.: *Chase, Chance, and Creativity: The Lucky Art of Novelty,* Updated Edition. Cambridge: MIT, 2003. xix, 245 p. Paper $18.95.

BAHRANI, ZAINAB: *The Graven Image: Representation in Babylonia and Assyria.* Philadelphia: Pennsylvania UP, 2003. x, 242 p. Cloth $49.95.

BAILEY, F. G.: *The Saving Lie: Truth and Method in the Social Sciences.* Philadelphia: Pennsylvania UP, 2003. xv, 213 p. Cloth $39.95.

BANHAM, GARY: *Kant's Practical Philosophy: From Critique to Doctrine.* New York: Palgrave Macmillan, 2003. ix, 270 p. Cloth $65.00.

BEISTEGUI, MIGUEL DE: *Thinking with Heidegger.* Bloomington: Indiana UP, 2003. xv, 211 p. Cloth $59.95, paper $24.95.

BENHABIB, SEYLA: *The Reluctant Modernism of Hannah Arendt,* New Edition. Lanham, MD: Rowman and Littlefield, 2003. lii, 261 p. Cloth $65.00, paper $24.95.

BENNETT, JONATHAN: *A Philosophical Guide to Conditionals.* New York: Oxford, 2003. xiii, 387 p. Cloth $80.00, paper $24.95.

BIELEFELDT, HEINER: *Symbolic Representation in Kant's Practical Philosophy.* New York: Cambridge, 2003. xiii, 202 p. Cloth $60.00.

BOWIE, ANDREW: *Introduction to German Philosophy: From Kant to Habermas.* Malden, MA: Polity, 2003. x, 291 p. Cloth $64.95, paper $27.95.

BOYLAN, MICHAEL, and JAMES A. DONAHUE: *Ethics across the Curriculum: A Practice-Based Approach.* Lanham, MD: Lexington, 2003. ix, 163 p. Cloth $55.00.

BREISACH, ERNST: *On the Future of History: The Postmodernist Challenge and Its Aftermath.* Chicago: University Press, 2003. ix, 243 p. Cloth $41.00, paper $16.00.

BRODY, BARUCH A.: *Taking Issue: Pluralism and Casuistry in Bioethics.* Washington, DC: Georgetown, 2003. x, 293 p. Cloth $49.95.

BUBNER, RÜDIGER: *The Innovations of Idealism.* New York: Cambridge, 2003. x, 274 p. Cloth $55.00.

BUDD, MALCOLM: *The Aesthetic Appreciation of Nature: Essays on the Aesthetics of Nature.* New York: Oxford, 2003. xii, 158 p. Cloth $29.95.

CASEBEER, WILLIAM D.: *Natural Ethical Facts: Evolution, Connectionism, and Moral Cognition.* Cambridge: MIT, 2003. x, 214 p. Cloth $35.00.

CHANDRA, SURESH. *Wittgenstein: New Perspectives.* New Delhi: Indian Council of Philosophical Research, 2002. viii, 157 p. Cloth.

CHIEN, ARNOLD J.: *On What We Mean.* Lanham, MD. UP of America, 2002. v, 229 p. Paper $38.00.

CHRISTMAN, JOHN: *Social and Political Philosophy.* New York: Routledge, 2002. xiii, 248 p. Paper $20.95.

CHURCHLAND, PATRICIA SMITH: *Brain-Wise: Studies in Neurophilosophy.* Cambridge: MIT, 2002. xii, 470 p. Paper $25.00.

CICOVACKI, PREDRAG: *Between Truth and Illusion: Kant at the Crossroads of Modernity*. Lanham, MD: Rowman and Littlefield, 2002. x, 229 p. Cloth $60.00, paper $21.95.

CLACK, BEVERLEY: *Sex and Death: A Reappraisal of Human Mortality*. Malden, MA: Polity, 2002. vi, 159 p. Cloth $59.95, paper $24.95.

CLADIS, MARK S.: *Public Vision, Private Lives: Rousseau, Religion, and 21st-century Democracy*. New York: Oxford, 2003. xvii, 298 p. Cloth $49.95.

CLARKE, D.S.: *Panpsychism and the Religious Attitude*. Albany: SUNY, 2003. x, 201 p. Cloth $65.50, paper $21.95.

CLOUGH, SHARYN: *Beyond Epistemology: A Pragmatic Approach to Feminist Science Studies*. Lanham, MD: Rowman and Littlefield, 2003. viii, 167 p. Cloth $65.00, paper $24.95.

COHEN, ELLIOT D.. *What Would Aristotle Do? Self-Control through the Power of Reason*. Amherst, NY: Prometheus, 2003. 251 p. Paper $21.00.

COLAPIETRO, VINCENT: *Fateful Shapes of Human Freedom: John William Miller and the Crises of Modernity*. Nashville, TN: Vanderbilt, 2003. xv, 323 p. Cloth $59.95, paper $27.95.

COLEBROOK, CLAIRE: *Irony in the Work of Philosophy*. Lincoln: Nebraska UP, 2002. xix, 332 p. Cloth $65.00.

CONLEY, JOHN J., S.J.: *The Suspicion of Virtue: Women Philosophers in Neoclassical France*. Ithaca, NY: Cornell, 2002. xi, 222. Cloth $39.95.

COOK, REBECCA J., BERNARD M. DICKENS, and MAHMOUD F. FATHALLA: *Reproductive Health and Human Rights*. New York: Oxford, 2003. xxvii, 554 p. Cloth $95.00, paper $29.95.

COOPER, DAVID E.: *The Measure of Things: Humanity, Humility, and Mystery*. New York: Oxford, 2002. ix, 372 p. Cloth $45.00, *World Philosophies: An Historical Introduction*, Second Edition. Malden, MA: Blackwell, 2003. viii, 571 p. Paper $84.95.

COOPER, WILLIAM S.: *The Evolution of Reason: Logic as a Branch of Biology*. New York: Cambridge, 2001. x, 226 p. Cloth $55.00.

COORS, MARILYN E.: *The Matrix: Charting an Ethics of Inheritable Genetic Modification*. Lanham, MD: Rowman and Littlefield, 2003. x, 161 p. Cloth $70.00, paper $23.95.

COPJEC, JOAN: *Imagine There's No Woman: Ethics and Sublimation*. Cambridge: MIT, 2003. vii, 261 p. Cloth $29.95.

CORFIELD, DAVID: *Towards a Philosophy of Real Mathematics*. New York: Cambridge, 2003. x, 288 p. Cloth $65.00.

CORLETT, J. ANGELO: *Race, Racism, and Reparations*. Ithaca, NY: Cornell, 2003. xi, 252 p. Cloth $47.50, paper $19.95.

COWAN, ROSEMARY: *Cornel West: The Politics of Redemption*. Malden, MA: Blackwell, 2003. viii, 216 p. Cloth $57.95, paper $22.95.

CRAIG, WILLIAM LANE, and WALTER SINNOTT-ARMSTRONG: *God? A Debate between a Christian and an Atheist*. New York: Oxford, 2004. xii, 156 p. Cloth $19.95, paper $15.95.

CRITCHLEY, SIMON: *On Humour*. New York: Routledge, 2002. ix, 132 p. Paper $12.95.

CRONIN, JOSEPH R., and OTTO R. PIECHOWSKI: *Returning Questions: A Dialogical Introduction to Philosophy.* Lanham, MD: UP of America. xi, 317 p. Paper $48.00.

CROWTHER, PAUL: *The Transhistorical Image: Philosophizing Art and Its History.* New York: Cambridge, 2002. x, 207 p. Cloth $55.00.

CUPITT, DON: *Is Nothing Sacred? The Non-realist Philosophy of Religion.* New York: Fordham, 2002. xxiii, 159 p. Cloth $40.00.

CURRIE, GREGORY, and IAN RAVENSCROFT: *Recreative Minds: Imagination in Philosophy and Psychology.* New York: Oxford, 2002. x, 233 p. Cloth $65.00, paper $21.95.

DALLMAYR, FRED R.: *G.W.F. Hegel: Modernity and Politics.* Lanham, MD: Rowman and Littlefield, 2002. xxi, 270 p. Cloth $65.00, paper $24.95.

DALTON, THOMAS C.: *Becoming John Dewey: Dilemmas of a Philosopher and Naturalist.* Bloomington: Indiana UP, 2002. xi, 375 p. Cloth $45.00.

DANIUS, SARA: *The Senses of Modernism: Technology, Perception, and Aesthetics.* Ithaca, NY: Cornell, 2002. xi, 297 p. Cloth $45.00, paper $19.95.

DANTO, ARTHUR C.: *The Abuse of Beauty: Aesthetics and the Concept of Art.* Chicago: Open Court, 2003. xxi, 167 p. Cloth $29.95, paper $19.95.

DARWALL, STEPHEN: *Welfare and Rational Care.* Princeton: University Press, 2002. xi, 135 p. Cloth $24.95.

DAVIES, OLIVER: *A Theology of Compassion: Metaphysics of Difference and the Renewal of Tradition.* Grand Rapids, MI: Eerdmans, 2003. xxii, 376 p. Paper $37.00.

DAVIES, STEPHEN: *Themes in the Philosophy of Music.* New York: Oxford, 2003. 283 p. Cloth $55.00.

DAVIS, WAYNE A.: *Meaning, Expression, and Thought.* New York Cambridge, 2003. xvii, 654 p. Cloth $90.00.

DE GEORGE, RICHARD T.: *The Ethics of Information Technology and Business.* Malden, MA: Blackwell, 2003. x, 289 p. Paper $24.95.

DELANCY, CRAIG: *Passionate Engines: What Emotions Reveal about Mind and Artificial Intelligence.* New York: Oxford, 2002. xiii, 254 p. Cloth $49.95, paper $26.00.

DESMOND, WILLIAM: *Art, Origins, Otherness: Between Philosophy and Art.* Albany: SUNY, 2003. xi, 306 p. Cloth $68.50, paper $22.95.

DEUTSCHER, PENELOPE: *A Politics of Impossible Difference: The Later Work of Luce Irigaray.* Ithaca, NY: Cornell, 2002. ix, 228 p. Cloth $39.95, paper $17.95.

DEVETTERE, RAYMOND J.: *Introduction to Virtue Ethics: Insights of the Ancient Greeks.* Washington, DC: Georgetown, 2002. ix, 195 p. Paper $21.95.

DICKENSON, DONNA: *Risk and Luck in Medical Ethics.* Malden, MA: Blackwell, 2003. x, 269 p. Paper $27.95.

DIETHE, CAROL: *Nietzsche's Sister and the Will to Power.* Urbana: Illinois UP, 2003. xiii, 218 p. Cloth $34.95.

DIXON, THOMAS: *From Passions to Emotions: The Creation of a Secular Psychological Category.* New York: Cambridge, 2003. x, 287 p. Cloth $60.00.

DORIS, JOHN M.: *Lack of Character: Moral Behavior.* New York: Cambridge, 2002. x, 272 p. Cloth $60.00.

DUDLEY, WILL: *Hegel, Nietzsche, and Philosophy: Thinking Freedom.* New York: Cambridge, 2002. xvii, 326 p. Cloth $60.00.

DUPRÉ, JOHN: *Humans and Other Animals.* New York: Oxford, 2002. vi, 272 p. Cloth $25.00.

DUSKA, RONALD F., and BRENDA SHAY DUSKA: *Accounting Ethics.* Malden, MA: Blackwell, 2003. li, 277 p. Paper $29.95.

DUVENAGE, PIETER: *Habermas and Aesthetics: The Limits of Communicative Reason.* Malden, MA: Polity, 2003. ix, 209 p. Cloth $64.95, paper $29.95.

EBERLE, CHRISTOPHER J.: *Religious Conviction in Liberal Politics.* New York: Cambridge, 2002. x, 405 p. Cloth $75.00, paper $28.00.

ECKSTEIN, JEROME: *On Meanings of Life: Their Nature and Origin.* Albany: SUNY, 2002. ix, 165 p. Cloth $54.50, paper $17.95.

ELBE, STEFAN: *Europe: A Nietzschean Perspective.* New York: Routledge, 2003. xiv, 168 p. Cloth $80.00.

ELLENBOGEN, SARA: *Wittgenstein's Account of Truth.* Albany: SUNY, 2003. xv, 148 p. Cloth $57.50, paper $18.95.

ELLIS, BRIAN: *The Philosophy of Nature: A Guide to the New Essentialism.* Montreal: McGill-Queen's, 2002. v, 185 p. Cloth $70.00, paper $22.95.

ELTON, MATTHEW: *Daniel Dennett: Reconciling Science and Our Self-conception.* Malden, MA: Polity, 2003. xiii, 296 p. Cloth $59.95, paper $25.95.

ENGEL, PASCAL: *Truth.* Montreal: McGill-Queen's, 2002. viii, 177 p. Cloth $70.00, paper $27.95.

EPSTEIN, RICHARD A.: *Skepticism and Freedom: A Modern Case for Classical Liberalism.* Chicago: University Press, 2003. viii, 311 p. Cloth $39.00.

FAAS, EKBERT: *The Genealogy of Aesthetics.* New York: Cambridge, 2002. xiii, 439 p. Cloth $65.00.

FAGG, LAWRENCE W.: *The Becoming of Time: Integrating Physical and Religious Time.* Durham, NC: Duke, 2003. xiii, 282 p. Paper $22.95.

FEINBERG, JOEL: *Problems at the Roots of Law: Essays in Legal and Political Theory.* New York: Oxford, 2003. viii, 220 p. Cloth $45.00.

FELDMAN, SEYMOUR: *Philosophy in a Time of Crisis: Don Isaac Abravanel: Defender of the Faith.* New York: Routledge, 2003. vi, 213 p. Cloth $75.00.

FENNER, DAVID E. W.: *Introducing Aesthetics.* Westport, CT: Praeger, 2003. viii, 171 p. Paper $29.95.

FESMIRE, STEVEN: *John Dewey and Moral Imagination: Pragmatism in Ethics.* Bloomington: Indiana UP, 2003. xii, 167 p. Cloth $49.95, paper $19.95.

FINE, KIT: *The Limits of Abstraction.* New York: Oxford, 2002. x, 203 p. Cloth $25 00.

FISTIOC, MIHAELA C.: *The Beautiful Shape of the Good: Platonic and Pythagorean Themes in Kant's Critique of the Power of Judgment.* New York: Routledge, 2002. xxiii, 160 p. Cloth $65.00.

FITTING, MELVIN: *Types, Tableaus, and Gödel's God.* Boston: Kluwer, 2002. xv, 180 p. Cloth $67.00.

FLEW, ANTONY: *Social Life and Moral Judgment.* New Brunswick, NJ: Transaction, 2003. ix, 179 p. Cloth $34.95.

FODOR, JERRY A., and ERNEST LEPORE: *The Compositionality Papers.* New York: Oxford, 2002 vi, 212 p. Cloth $55.00.

FOGELIN, ROBERT J.: *A Defense of Hume on Miracles.* Princeton: University Press,

2003. xii, 101 p. Cloth $24.95; *Walking the Tightrope of Reason: The Precarious Life of a Rational Animal.* New York: Oxford, 2003. xii, 203 p. Cloth $22.00.

FOOT, PHILIPPA: *Moral Dilemmas.* New York: Oxford, 2002. 218 p. Cloth $55.00, paper $19.95.

FORSTER, THOMAS: *Logic, Induction, and Sets.* New York: Cambridge, 2003. x, 234 p. Cloth $75.00, paper $27.00.

FRANKLIN, ALLAN: *Selectivity and Discord: Two Problems of Experiment.* Pittsburgh: University Press, 2002. 290 p. Cloth $37.50.

FRIEDMAN, MARILYN: *Autonomy, Gender, Politics.* New York: Oxford, 2003. ix, 248 p. Cloth $49 95, paper $19.95.

FRIERSON, PATRICK R.: *Freedom and Anthropology in Kant's Moral Philosophy.* New York: Cambridge, 2003. x, 211 p. Cloth $55.00.

GANS, CHAIM: *The Limits of Nationalism.* New York: Cambridge, 2003. ix, 192 p. Cloth $60.00, paper $22.00.

GARRETT, AARON V.: *Meaning in Spinoza's Method.* New York: Cambridge, 2003. xii, 240 p. Cloth $60.00.

GASCHÉ, RODOLPHE: *The Idea of Form: Rethinking Kant's Aesthetics.* Stanford: University Press, 2003. x, 259 p. Cloth $55.00, paper $24.95.

GAUKER, CHRISTOPHER: *Words without Meaning.* Cambridge: MIT, 2003. xi, 299 p. Paper $25.00.

GELVEN, MICHAEL: *What Happens to Us When We Think: Transformation and Reality.* Albany: SUNY, 2003. 149 p. Cloth $49.50, paper $16.95.

GERSON, LLOYD P.: *Knowing Persons: A Study in Plato.* New York: Oxford, 2003. x, 308 p. Cloth $45.00.

GIAQUINTO, MARCUS: *The Search for Certainty: A Philosophical Account of Foundations of Mathematics.* New York: Oxford, 2002. x, 286 p. Cloth $52.00.

GIBBARD, ALLAN: *Thinking How to Live.* Cambridge: Harvard, 2003. xv, 203 p. Cloth $45.00.

GILBERT, PAUL: *New Terror, New Wars.* Washington DC: Georgetown, 2003. vii, 164 p. Paper $24.95.

GILL, FRANCES, E.: *The Moral Benefit to Punishment: Self-determination as a Goal of Correctional Counseling.* Lanham, MD: Lexington, 2003. xiii, 157 p. Cloth $60.00.

GINS, MADELINE, and ARAKAWA: *Architectural Body.* Tuscaloosa: Alabama UP, 2002. xxii, 102 p. Cloth $45.00, paper $19.95.

GLOCK, HANS-JOHANN: *Quine and Davidson on Language, Thought, and Reality.* New York: Cambridge, 2003. xvi, 311 p. Cloth $60.00.

GODFREY-SMITH, PETER: *Theory and Reality: An Introduction to the Philosophy of Science.* Chicago: University Press, 2003. xiii, 271 p. Paper $24.00.

GORDON, PETER ELI: *Rosenzweig and Heidegger: Between Judaism and German Philosophy.* Berkeley: California UP, 2003. xxix, 331 p. Cloth $65.00.

GOURGOURIS, STATHIS: *Does Literature Think? Literature as Theory for an Anti-mythical Era.* Stanford: University Press, 2003. xxvi, 391 p. Cloth $60.00, paper $22.95.

GRACIA, JORGE J.E.: *Old Wine in New Skins.* Milwaukee: Marquette, 2003. 140 p. Cloth $15.00.

GRIMMETT, GEOFFREY, and DAVID STIRZAKER: *Probability and Random Process*, Third Edition. New York: Oxford, 2001. xii, 596 p. Paper $54.95.

GRUNEBAUM, JAMES O.: *Friendship: Liberty, Equality, and Utility*. Albany: SUNY, 2003. ix, 192 p. Cloth $57.50, paper $18.95.

GUPTA, BINA: *CIT: Consciousness*. New York: Oxford, 2003. xi, 203 p. Cloth $35.00.

GUPTA, CHHANDA: *Realism versus Realism*. Lanham, MD: Rowman and Littlefield, 2003. xvi, 161 p. Cloth $60.00, paper $24.95.

GUTMANN, AMY: *Identity in Democracy*. Princeton: University Press, 2003. ix, 246 p. Cloth $27.95.

HAACK, SUSAN: *Defending Science—Within Reason: Between Scientism and Cynicism*. Amherst, NY: Prometheus, 2003. 411 p. Cloth $28.00.

HAJI, ISHTIYAQUE: *Deontic Morality and Control*. New York: Cambridge, 2002. xiv, 288 p. Cloth $65.00.

HALLER, STEPHEN F.: *Apocalypse Soon? Wagering on Warnings of Global Catastrophe*. Montreal: McGill-Queen's, 2002. xii, 185 p. Cloth $65.00, paper $19.95.

HALLIWELL, STEPHEN: *The Aesthetics of Mimesis: Ancient Texts and Modern Problems*. Princeton: University Press, 2002. ix, 424 p. Cloth $65.00, paper $24.95.

HAMILTON, LAWRENCE A.: *The Political Philosophy of Needs*. New York: Cambridge, 2003. x, 209 p. Cloth $60.00.

HANBY, MICHAEL: *Augustine and Modernity*. New York: Routledge, 2003. xii, 292 p. Cloth $90.00.

HARDIN, RUSSELL: *Indeterminacy and Society*. Princeton: University Press, 2003. xii, 166 p. Cloth $29.95.

HARMAN, GRAHAM: *Tool-being: Heidegger and the Metaphysics of Objects*. Chicago: Open Court, 2002. viii, 331 p. Paper $29.95.

HARRISON, ROBERT POGUE: *The Dominion of the Dead*. Chicago: University Press, 2003. xiii, 208 p. Cloth $22.50.

HARRISON, ROSS: *Hobbes, Locke, and Confusion's Masterpiece: An Examination of Seventeenth-century Political Philosophy*. New York: Cambridge, 2003. v, 281 p. Cloth $65.00, paper $23.00.

HARTE, VERITY: *Plato on Parts and Wholes*. New York: Oxford, 2002. x, 311 p. Cloth $45.00.

HARTLE, ANN: *Michel de Montaigne: Accidental Philosopher*. New York: Cambridge, 2003. viii, 303 p. Cloth $60.00.

HARTLEY, GEORGE: *The Abyss of Representation: Marxism and the Postmodern Sublime*. Durham, NC: Duke, 2003. xiii, 338 p. Cloth $69.95, paper $22.95.

HARVEY, IRENE E.: *Labyrinths of Exemplarity: At the Limits of Deconstruction*. Albany: SUNY, 2002. x, 281 p. Cloth $81.50, paper $27.95.

HEIL, JOHN: *From an Ontological Point of View*. New York: Oxford, 2003. xv, 267 p. Cloth $45.00.

HENN, MARTIN J.: *Parmenides of Elea: A Verse Translation with Interpretative Essays and Commentary to the Text*. Westport, CT: Praeger, 2003. 149 p. Cloth $59.95.

HENRICH, DIETER: *Between Kant and Hegel: Lectures on German Idealism*. DAVID S. PACINI, ed. Cambridge: Harvard, 2003. liv, 841 p. Cloth $55.00.

HERBERT, GARY B.: *A Philosophical History of Rights*. New Brunswick, NJ: Transaction, 2002. xviii, 362 p. Cloth $49.95.

HIGGS, ERIC: *Nature by Design: People, Natural Process, and Ecological Restoration.* Cambridge: MIT, 2003. xv, 341 p. Paper $27.95.

HILL, CHRISTOPHER S.: *Thought and World: An Austere Portrayal of Truth, Reference, and Semantic Correspondence.* New York: Cambridge, 2002. xii, 154 p. Cloth $65.00, paper $23.00.

HILL, R. KEVIN: *Nietzsche's Critiques: The Kantian Foundations of His Thought.* New York: Oxford, 2003. xvi, 242 p. Cloth $45.00.

HOLLAND, STEPHEN: *Bioethics: A Philosophical Introduction.* Malden, MA: Polity, 2003. vi, 232 p. Cloth $59.95, paper $24 95.

HONDERICH, TED: *How Free Are You? The Determinism Problem,* Second Edition New York: Oxford, 2002. 176 p. Paper $19.95.

HOWIE, DAVID: *Interpreting Probability: Controversies and Developments in the Early Twentieth Century.* New York: Cambridge, 2002. xi, 262 p. Cloth $60.00.

HOWIE, GILLIAN: *Deleuze and Spinoza: Aura of Expressionism.* New York: Palgrave MacMillan, 2002. ix, 242 p. Cloth $68.00.

HURLEY, S.L.: *Justice, Luck, and Language.* Cambridge: Harvard, 2003. viii, 341 p. Cloth $55.00.

HUSSEIN, ABDIRAHMAN A.: *Edward Said: Criticism and Society.* New York: Verso, 2002. ix, 339 p. Cloth $25.00.

HUYSSEN, ANDREAS: *Present Pasts: Urban Palimpsests and the Politics of Memory.* Stanford: University Press, 2003. xi, 177 p. Cloth $49.50, paper $19.95.

HYDER, DAVID: *The Mechanics of Meaning: Propositional Content and the Logical Space of Wittgenstein's Tractatus.* New York: de Gruyter, 2002. ix, 229 p. Cloth $81.60.

JACOBS, JO ELLEN: *The Voice of Harriet Taylor Mill.* Bloomington: Indiana UP, 2002. xxviii, 270 p. Cloth $39.95.

JACOBS, JONATHAN: *Dimensions of Moral Theory: An Introduction to Metaethics and Moral Psychology.* Malden, MA: Blackwell, 2002. xii, 175 p. Cloth $22.95.

JACQUETTE, DALE: *Ontology.* Montreal: McGill-Queen's, 2002 xiv, 348 p. Cloth $70.00, paper $27.95.

JAMIESON, DALE: *Morality's Progress.* New York: Oxford, 2002. xiv, 380 p. Cloth $72.00, paper $24.95.

JOHNSTON, GEORGE F.: *Abortion from the Religious and Moral Perspective: An Annotated Bibliography.* Westport, CT: Praeger, 2003. xiii, 248 p. Cloth $79.95.

KEENAN, ALAN: *Democracy in Question: Democratic Openness in a Time of Political Closure.* Stanford: University Press, 2003. xi, 238 p. Cloth $49.50, paper $19.95.

KEKES, JOHN: *The Art of Life.* Ithaca, NY: Cornell, 2002. xii, 267 p. Cloth $29.95.

KELLY, MICHAEL: *Iconoclasm in Aesthetics.* New York: Cambridge, 2003. xiii, 222 p. Cloth $60.00.

KERR, FERGUS: *After Aquinas: Versions of Thomism.* Malden, MA: Blackwell, 2002. viii, 254 p. Paper $26.95.

KESTENBAUM, VICTOR: *The Grace and the Severity of the Ideal: John Dewey and the Transcendent.* Chicago: University Press, 2002. 261 p. Paper $20.00.

KITCHER, PHILIP: *In Mendel's Mirror: Philosophical Reflections on Biology.* New York: Oxford, 2003. xv, 385 p. Paper $75.00, paper $35.00.

INDEX TO VOLUME C

INDEX OF SUBJECTS DISCUSSED IN THE JOURNAL

Aesthetics

Ridley, Aaron: Against Musical Ontology. 100 (4). 203–20 '03

Andreasen, Robin

Glasgow, Joshua M: On the New Biology of Race 100 (9): 456–75 '03

Bayesian Theory

Arntzenius, Frank Some Problems for Conditionalization and Reflection. 100 (7): 356–70 '03

Godfrey-Smith, Peter: Goodman's Problem and Scientific Methodology. 100 (11): 573–90 '03

Peterson, Martin: From Consequentialism to Utilitarianism. 100 (8). 403–15 '03

Belief

Arntzenius, Frank: Some Problems for Conditionalization and Reflection. 100 (7). 356–70 '03

Cantwell, John: On the Foundations of Pragmatic Arguments. 100 (8): 383–403 '03

Biology

Glasgow, Joshua M.: On the New Biology of Race. 100 (9): 456–75 '03

Hardimon, Michael O.: The Ordinary Concept of Race 100 (9): 437–55 '03

Chalmers, David

Braddon-Mitchell, David: Qualia and Analytical Conditionals. 100 (3): 111–35 '03

Choice

Cantwell, John On the Foundations of Pragmatic Arguments. 100 (8): 383–403 '03

Christofidou, Andrea

Coliva, Annalisa: The First Person: Error through Misidentification, the Split between Speaker's and Semantic Reference, and the Real Guarantee. 100 (8). 416–31 '03

Clapp, Lenny

Gillett, Carl: The Metaphysics of Realization, Multiple Realizability, and the Special Sciences. 100 (11): 591–603 '03

Cognitive Science

Eliasmith, Chris: Moving beyond Metaphors: Understanding the Mind for What It Is 100 (10): 493–520 '03

Color

Hardin, C L.: A Spectral Reflectance Doth Not a Color Make 100 (4): 191–202 '03

Conditionalization

Arntzenius, Frank: Some Problems for Conditionalization and Reflection. 100 (7): 356–70 '03

Consequentialism

Peterson, Martin: From Consequentialism to Utilitarianism. 100 (8): 403–15 '03

Phillips, David. Thomson and the Semantic Argument against Consequentialism. 100 (9): 475–86 '03

Schapiro, Tamar: Compliance, Complicity, and the Nature of Nonideal Conditions. 100 (7): 329–55 '03

Counterfactuals

Godfrey-Smith, Peter Goodman's Problem and Scientific Methodology. 100 (11): 573–90 '03

Tooley, Michael: The Stalnaker–Lewis Approach to Counterfactuals. 100 (7): 371–77 '03

Counterpart Theory

Merricks, Trenton: The End of Counterpart Theory. 100 (10): 521–49 '03

Davidson, Donald

Dummett, Michael: The Dewey Lectures 2002: Truth and the Past. 100 (1): 5–53 '03

Decision Theory

Cantwell, John: On the Foundations of Pragmatic Arguments. **100** (8): 383–403 '03

Peterson, Martin. From Consequentialism to Utilitarianism. **100** (8): 403–15 '03

Desire

Bratman, Michael E.: A Desire of One's Own **100** (5): 221–42 '03

Essentialism

Hardimon, Michael O.: The Ordinary Concept of Race **100** (9): 437–55 '03

Ethics

Gibbard, Allan Reasons Thin and Thick. **100** (6): 288–304 '03

Peterson, Martin: From Consequentialism to Utilitarianism. **100** (8): 403–15 '03

Phillips, David: Thomson and the Semantic Argument against Consequentialism. **100** (9): 475–86 '03

Schapiro, Tamar: Compliance, Complicity, and the Nature of Nonideal Conditions. **100** (7). 329–55 '03

Taylor, Charles. Ethics and Ontology **100** (6): 305–20 '03

Explanation

Stanford, P. Kyle: Pyrrhic Victories for Scientific Realism. **100** (11): 553–72 '03

Explanation, Scientific

Derden, James: A Different Conception of Scientific Realism: The Case for the Missing Explanada. **100** (5): 243–67 '03

Fodor, Jerry

Gillett, Carl: The Metaphysics of Realization, Multiple Realizability, and the Special Sciences **100** (11): 591–603 '03

Frankfurt, Harry

Bratman, Michael E A Desire of One's Own **100** (5): 221–42 '03

Free Will

Bratman, Michael E.: A Desire of One's Own **100** (5): 221–42 '03

Frege, Gottlob

Fine, Kit: The Role of Variables **100** (12): 605–31 '03

Good

Phillips, David. Thomson and the Semantic Argument against Consequentialism. **100** (9): 475–86 '03

Goodman, Nelson

Godfrey–Smith, Peter. Goodman's Problem and Scientific Methodology. **100** (11). 573–90 '03

Ridley, Aaron: Against Musical Ontology. **100** (4) 203–20 '03

Grice, H.P.

O'Rourke, Michael: The Scope Argument. **100** (3) 156–57 '03

Grue

Godfrey–Smith, Peter. Goodman's Problem and Scientific Methodology. **100** (11): 573–90 '03

Heller, Mark

Merricks, Trenton: The End of Counterpart Theory. **100** (10): 521–49 '03

Identity, Personal

Coliva, Annalisa: The First Person: Error through Misidentification, the Split between Speaker's and Semantic Reference, and the Real Guarantee. **100** (8): 416–31 '03

Induction

Godfrey–Smith, Peter: Goodman's Problem and Scientific Methodology. **100** (11): 573–90 '03

Jackson, Frank

Godfrey–Smith, Peter: Goodman's Problem and Scientific Methodology. **100** (11): 573–90 '03

Justification

Gibbard, Allan: Reasons Thin and Thick. **100** (6): 288–304 '03

Kant, Immanuel

Schapiro, Tamar: Compliance, Complicity, and the Nature of Nonideal Conditions. **100** (7): 329–55 '03

Knowledge, Theory of

Gibbard, Allan Reasons Thin and Thick. **100** (6): 288–304 '03

Rumfitt, Ian: Savoir Faire. **100** (3): 158–66 '03

Strevens, Michael: The Role of the Priority Rule in Science. **100** (2). 83–79 '03

Language, Natural

O'Rourke, Michael The Scope Argument. **100** (3): 136–57 '03

Language, Philosophy of

Fine, Kit: The Role of Variables. **100** (12): 605–31 '03

Lewis, David

Merricks, Trenton. The End of Counterpart Theory **100** (10): 521–49 '03

Tooley, Michael The Stalnaker–Lewis Approach to Counterfactuals. **100** (7). 371–77 '03

Locke, John

Coliva, Annalisa. The First Person: Error through Misidentification, the Split between Speaker's and Semantic Reference, and the Real Guarantee. **100** (8): 416–31 '03

Logic

Fine, Kit· The Role of Variables. **100** (12): 605–31 '03

O'Rourke, Michael: The Scope Argument. **100** (3): 136–57 '03

Tooley, Michael· The Stalnaker–Lewis Approach to Counterfactuals **100** (7)· 371–77 '03

Mathematics, Philosophy of

Derden, James A Different Conception of Scientific Realism. The Case for the Missing Explanada. **100** (5)· 243–67 '03

Fine, Kit: The Role of Variables. **100** (12): 605–31 '03

McDowell, John

Coliva, Annalisa· The First Person: Error through Misidentification, the Split between Speaker's and Semantic Reference, and the Real Guarantee. **100** (8): 416–31 '03

Taylor, Charles: Ethics and Ontology. **100** (6) 305–20 '03

Meaning, Theory of

Dummett, Michael: The Dewey Lectures 2002: Truth and the Past. **100** (1). 5–53 '03

Measurement

Fine, Arthur: Quantum Life: Interaction, Entanglement, and Separation. By Eric Winsberg and Arthur Fine. **100** (2): 80–97 '03

Winsberg, Eric: Quantum Life: Interaction, Entanglement, and Separation. By Eric Winsberg and Arthur Fine. **100** (2). 80–97 '03

Memo to Authors

100 (1): 54 '03

Metaphysics

Dummett, Michael: The Dewey Lectures 2002. Truth and the Past. **100** (1): 5–53 '03

Gillett, Carl: The Metaphysics of Realization, Multiple Realizability, and the Special Sciences. **100** (11): 591–603 '03

Merricks, Trenton: The End of Counterpart Theory. **100** (10): 521–49 '03

Mind, Philosophy of

Braddon–Mitchell, David: Qualia and Analytical Conditionals **100** (3): 111–35 '03

Eliasmith, Chris. Moving beyond Metaphors Understanding the Mind for What It Is. **100** (10): 493–520 '03

Gillett, Carl: The Metaphysics of Realization, Multiple Realizability, and the Special Sciences **100** (11): 591–603 '03

Moore, G.E.

Phillips, David: Thomson and the Semantic Argument against Consequentialism. **100** (9): 475–86 '03

Multiple Realizability

Gillett, Carl The Metaphysics of Realization, Multiple Realizability, and the Special Sciences **100** (11): 591–603 '03

Music

Ridley, Aaron: Against Musical Ontology **100** (4): 203–20 '03

Nationalism, Ethical

Taylor, Charles: Ethics and Ontology. **100** (6): 305–20 '03

Neuroscience

Elliasmith, Chris Moving beyond Metaphors: Understanding the Mind for What It Is. **100** (10): 493–520 '03

New Books

100 (8): 435–36 '03
100 (12). 644–50 '03

New Books: Anthologies

100 (7) 378–82 '03
100 (8): 432–34 '03

New Books: Translations

100 (5): 273–74 '03
100 (6): 326–27 '03

Notes and News

100 (2). 109–10 '03
100 (10). 550 '03
100 (12): 642–43 '03

Ontology

Ridley, Aaron: Against Musical Ontology. **100** (4): 203–20 '03

Taylor, Charles: Ethics and Ontology. **100** (6): 305–20 '03

Perception

Hardin, C.L. A Spectral Reflectance Doth Not a Color Make **100** (4): 191–202 '03

Pessimistic Induction

Stanford, P. Kyle: Pyrrhic Victories for Scientific Realism. **100** (11). 553–72 '03

Physicalism

Braddon-Mitchell, David: Qualia and Analytical Conditionals. **100** (3). 111–35 '03

Gillett, Carl: The Metaphysics of Realization, Multiple Realizability, and the Special Sciences. **100** (11): 591–603 '03

Possible Worlds

Merricks, Trenton The End of Counterpart Theory. **100** (10): 521–49 '03

Pragmatism

Price, Huw. Truth as Convenient Friction **100** (4): 167–90 '03

Priority Rule

Strevens, Michael The Role of the Priority Rule in Science. **100** (2): 55–79 '03

Probability

Fine, Arthur: Quantum Life: Interaction, Entanglement, and Separation By Eric Winsberg and Arthur Fine **100** (2): 80–97 '03

Winsberg, Eric. Quantum Life: Interaction, Entanglement, and Separation. By Eric Winsberg and Arthur Fine. **100** (2): 80–97 '03

Psychology, Philosophy of

Elliasmith, Chris Moving beyond Metaphors. Understanding the Mind for What It Is. **100** (10). 493–520 '03

Gillett, Carl. The Metaphysics of Realization, Multiple Realizability, and the Special Sciences. **100** (11): 591–603 '03

Putnam, Hilary

Gillett, Carl. The Metaphysics of Realization, Multiple Realizability, and the Special Sciences **100** (11): 591–603 '03

Qualia

Braddon-Mitchell, David. Qualia and Analytical Conditionals. **100** (3): 111–35 '03

Quantum Mechanics

Fine, Arthur: Quantum Life. Interaction, Entanglement, and Separation. By Eric Winsberg and Arthur Fine. **100** (2) 80–97 '03

Winsberg, Eric: Quantum Life: Interaction, Entanglement, and Separation. By Eric Winsberg and Arthur Fine **100** (2) 80–97 '03

Race

Glasgow, Joshua M.. On the New Biology of Race. **100** (9): 456–75 '03

Hardimon, Michael O The Ordinary Concept of Race. **100** (9): 437–55 '03

Rationality

Arntzenius, Frank. Some Problems for Conditionalization and Reflection **100** (7): 356–70 '03

Cantwell, John: On the Foundations of Pragmatic Arguments. **100** (8): 383–403 '03

Gibbard, Allan: Reasons Thin and Thick. **100** (6): 288–304 '03

Schapiro, Tamar: Compliance, Complicity, and the Nature of Nonideal Conditions. **100** (7): 329–55 '03

Rawls, John

Schapiro, Tamar: Compliance, Complicity, and the Nature of Nonideal Conditions. **100** (7): 329–55 '03

Realism

Derden, James: A Different Conception of Scientific Realism: The Case for the Missing Explanada. **100** (5): 243–67 '03

Hardin, C.L.: A Spectral Reflectance Doth Not a Color Make. **100** (4): 191–202 '03

Stanford, P. Kyle: Pyrrhic Victories for Scientific Realism **100** (11) 553–72 '03

Realism, Modal

Merricks, Trenton: The End of Counterpart Theory. **100** (10): 521–49 '03

Reference

Dummett, Michael The Dewey Lectures 2002: Truth and the Past. **100** (1): 5–53 '03

Right

Phillips, David: Thomson and the Semantic Argument against Consequentialism. **100** (9): 475–86 '03

Rorty, Richard

Price, Huw: Truth as Convenient Friction **100** (4) 167–90 '03

Rosenberg, Alex

Gillett, Carl The Metaphysics of Realization, Multiple Realizability, and the Special Sciences. **100** (11) 591–603 '03

Rovane, Carol

Coliva, Annalisa The First Person: Error through Misidentification, the Split between Speaker's and Semantic Reference, and the Real Guarantee. **100** (8): 416–31 '03

Ryle, Gilbert

Rumfitt, Ian: Savoir Faire. **100** (3): 158–66 '03

Schapiro, Lawrence

Gillett, Carl: The Metaphysics of Realization, Multiple Realizability, and the Special Sciences. **100** (11): 591–603 '03

Science, Philosophy of

Derden, James: A Different Conception of Scientific Realism: The Case for the Missing Explanada. **100** (5): 243–67 '03

Fine, Arthur: Quantum Life: Interaction, Entanglement, and Separation By Eric Winsberg and Arthur Fine. **100** (2) 80–97 '03

Gillett, Carl: The Metaphysics of Realization, Multiple Realizability, and the Special Sciences. **100** (11). 591–603 '03

Glasgow, Joshua M : On the New Biology of Race. **100** (9). 456–75 '03

Godfrey-Smith, Peter: Goodman's Problem and Scientific Methodology. **100** (11): 573–90 '03

Hardimon, Michael O.: The Ordinary Concept of Race. **100** (9): 437–55 '03

Hardin, C.L.: A Spectral Reflectance Doth Not a Color Make. **100** (4): 191–202 '03

Stanford, P. Kyle: Pyrrhic Victories for Scientific Realism. **100** (11). 553–72 '03

Strevens, Michael: The Role of the Priority Rule in Science. **100** (2). 55–79 '03

Winsberg, Eric: Quantum Life: Interaction, Entanglement, and Separation. **100** (2): 80–97 '03

Self-reference

Coliva, Annalisa: The First Person: Error through Misidentification, the Split between Speaker's and Semantic Reference, and the Real Guarantee. **100** (8): 416–31 '03

Semantics

Dummett, Michael: The Dewey Lectures 2002: Truth and the Past **100** (1). 5–53 '03

Fine, Kit: The Role of Variables. **100** (12): 605–31 '03

O'Rourke, Michael. The Scope Argument. **100** (3): 156–57 '03

Social Theory

Glasgow, Joshua M.: On the New Biology of Race. **100** (9): 456–75 '03

Hardimon, Michael O.· The Ordinary Concept of Race **100** (9): 437–55 '03

Stalnaker, Robert

Tooley, Michael: The Stalnaker–Lewis Approach to Counterfactuals. **100** (7): 371–77 '03

Stanley, Jason

Rumfitt, Ian: Savoir Faire **100** (3). 158–66 '03

Tarski, Alfred

Fine, Kit: The Role of Variables **100** (12): 605–31 '03

Theories

Derden, James: A Different Conception of Scientific Realism: The Case for the Musung Explanada. **100** (5). 243–67 '03

Stanford, P. Kyle· Pyrrhic Victories for Scientific Realism. **100** (11): 553–72 '03

Thompson, Judith Jarvis

Phillips, David: Thomson and the Semantic Argument against Consequentialism. **100** (9)· 475–86 '03

Time

Dummett, Michael The Dewey Lectures 2002· Truth and the Past. **100** (1): 5–53 '03

Truth

Price, Huw: Truth as Convenient Friction. **100** (4)· 167–90 '03

Tye, Michael

Hardin, C.L: A Spectral Reflectance Doth Not a Color Make. **100** (4)· 191–202 '03

Utilitarianism

Peterson, Martin: From Consequentialism to Utilitarianism. **100** (8). 403–15 '03

Schapiro, Tamar Compliance, Complicity, and the Nature of Nonideal Conditions. **100** (7) 329–55 '03

Value, Theory of

Bratman, Michael E.: A Desire of One's Own. **100** (5)· 221–42 '03

van Fraassen, Bas

Arntzenius, Frank Some Problems for Conditionalization and Reflection **100** (7). 356–70 '03

Derden, James. A Different Conception of Scientific Realism: The Case for the Musung Explanada. **100** (5) 243–67 '03

Variables

Fine, Kit· The Role of Variables **100** (12): 605–31 '03

Vision

Hardin, C.L: A Spectral Reflectance Doth Not a Color Make. **100** (4): 191–202 '03

Williams, Bernard

Gibbard, Allan: Reasons Thin and Thick. **100** (6): 288–304 '03

Scanlon, T.M.: Thickness and Theory. **100** (6): 275–87 '03

Taylor, Charles: Ethics and Ontology. **100** (6)· 305–20 '03

Williamson, Timothy

Rumfitt, Ian: Savoir Faire. **100** (3). 158–66 '03

Wittgenstein, Ludwig

Dummett, Michael: The Dewey Lectures 2002 Truth and the Past. **100** (1) 5–53 '03

Zombies

Braddon–Mitchell, David: Qualia and Analytical Conditionals **100** (3): 111–35 '03

INDEX OF AUTHORS

Including Contributing Authors and Reviewers

Armstrong, Frank, Some Problems for Conditionalization and Reflection. **100** (7): 356–70 '03

Braddon-Mitchell, David, Qualia and Analytical Conditionals. **100** (3): 111–35 '03

Bratman, Michael E., A Desire of One's Own. **100** (5): 221–42 '03

Cantwell, John, On the Foundations of Pragmatic Arguments. **100** (8): 383–403 '03

Chatterjee, Partha, Review of Jeremy Waldron: *God, Locke, and Equality: Christian Foundations in Locke's Political Thought.* **100** (12). 638–41 '03

Colliva, Annalisa, The First Person: Error through Misidentification, the Split between Speaker's and Semantic Reference, and the Real Guarantee. **100** (8): 416–31 '03

Derden, James, A Different Conception of Scientific Realism: The Case for the Missing Explanada. **100** (5): 243–67 '03

Dummett, Michael, The Dewey Lectures 2002: Truth and the Past. **100** (1): 5–53 '03

Elizamith, Chris, Moving beyond Metaphors: Understanding the Mind for What It Is **100** (10): 493–520 '03

Fine, Arthur, Quantum Life. Interaction, Entanglement, and Separation. By Eric Winsberg and Arthur Fine. **100** (2): 80–97 '03

Fine, Kit, The Role of Variables. **100** (12): 605–31 '03

Fischer, John Martin, Review of Daniel C. Dennett: *Freedom Evolves.* **100** (12): 632–37 '03

Gibbard, Allan, Reasons Thin and Thick **100** (6): 288–304 '03

Gillett, Carl, The Metaphysics of Realization, Multiple Realizability, and the Special Sciences **100** (11): 591–603 '03

Glasgow, Joshua M., On the New Biology of Race **100** (9). 456–75 '03

Godfrey-Smith, Peter, Goodman's Problem and Scientific Methodology **100** (11): 573–90 '03

Hardimon, Michael O., The Ordinary Concept of Race. **100** (9): 437–55 '03

Hardin, C.L., A Spectral Reflectance Doth Not a Color Make. **100** (4): 191–202 '03

Jackson, Frank, Review of Donald Davidson: *Subjective, Intersubjective, Objective.* **100** (6): 321–25 '03

Longuenesse, Béatrice, Review of Henry E. Allison: *Kant's Theory of Taste: A Reading of the Critique of Aesthetic Judgment.* **100** (9): 487–92 '03

Merricks, Trenton, The End of Counterpart Theory. **100** (10): 521–49 '03

O'Rourke, Michael, The Scope Argument. **100** (3): 136–57 '03

Peterson, Martin, From Consequentialism to Utilitarianism. **100** (8). 403–15 '03

Phillips, David, Thomson and the Semantic Argument against Consequentialism **100** (9): 475–86 '03

Price, Huw, Truth as Convenient Friction. **100** (4): 167–90 '03

Priest, Graham, Review of Timothy Williamson: *Knowledge and Its Limits.* **100** (5): 268–72 '03

Ridley, Aaron, Against Musical Ontology. **100** (4) 203–20 '03

Rumfitt, Ian, Savoir Faire, **100** (5) 158–66 '03

Scanlon, T.M., Thickness and Theory **100** (6): 275–87 '03

Schapiro, Tamar, Compliance, Complicity, and the Nature of Nonideal Conditions. **100** (7): 329–55 '03

Stanford, P. Kyle, Pyrrhic Victories for Scientific Realism. **100** (11). 553–72 '03

Strevens, Michael, The Role of the Priority Rule in Science **100** (2): 55–79 '03

Taschek, William W., Review of Robert Stalnaker: *Context and Content.* **100** (2): 98–108 '03

Taylor, Charles, Ethics and Ontology. **100** (6): 305–20 '03

Tooley, Michael, The Stalnaker-Lewis Approach to Counterfactuals. **100** (7): 371–77 '03

Winsberg, Eric, Quantum Life: Interaction, Entanglement, and Separation. By Eric Winsberg and Arthur Fine. **100** (2): 80–97 '03

INDEX OF BOOKS REVIEWED

Listed by Author of Item Reviewed

Allison, Henry E.: *Kant's Theory of Taste: A Reading of the Critique of Aesthetic Judgment*—Béatrice Longuenesse. **100** (9) 487–92 '03

Davidson, Donald: *Subjective, Intersubjective, Objective*—Frank Jackson. **100** (6): 321–25 '03

Dennett, Daniel C.: *Freedom Evolves*—John Martin Fischer. **100** (12): 632–37 '03

Stalnaker, Robert: *Context and Content*—William W. Taschek. **100** (2): 98–108 '03

Waldron, Jeremy: *God, Locke, and Equality. Christian Foundations in Locke's Political Thought*—Partha Chatterjee. **100** (12). 638–41 '03

Williamson, Timothy: *Knowledge and Its Limits*—Graham Priest. **100** (5). 268–72 '03

United States Postal Service

Statement of Ownership, Management, and Circulation

1. Publication Title	2. Publication Number	3. Filing Date
THE JOURNAL OF PHILOSOPHY	0 0 2 2 - 3 6 2 X	10/10/03
4. Issue Frequency	5. Number of Issues Published Annually	6. Annual Subscription Price
MONTHLY	12	Individual:$35.00 Institution:$75.00

7. Complete Mailing Address of Known Office of Publication (Not printer) (Street, city, county, state, and ZIP+4)
The Journal of Philosophy, Columbia University
1150 Amsterdam Avenue, MC 4972
New York NY 10027

Contact Person: Pam Ward
Telephone: 212 666 4419

8. Complete Mailing Address of Headquarters or General Business Office of Publisher (Not printer)

SAME

9. Full Names and Complete Mailing Addresses of Publisher, Editor, and Managing Editor (Do not leave blank)

Publisher (Name and complete mailing address)
The Journal of Philosophy, Inc
1150 Amsterdam Avenue, MC 4972
New York NY 10027

Editor (Name and complete mailing address)
Arthur Danto, President
The Journal of Philosophy, Inc , 1150 Amsterdam Avenue, MC 4972
New York NY 10027

Managing Editor (Name and complete mailing address)
John Searle, Managing Editor
The Journal of Philosophy, Inc , 1150 Amsterdam Avenue, MC 4972
New York NY 10027

10. Owner (If owned by a corporation, give the name and address of the corporation immediately followed by the names and addresses of all stockholders owning or holding 1 percent or more of the total amount of stock. If not owned by a corporation, give the names and addresses of the individual owners. If owned by a partnership or other unincorporated firm, give its name and address as well as those of each individual owner. If the publication is published by a nonprofit organization, give its name and address.)

Full Name	Complete Mailing Address
The Journal of Philosophy, Inc	Columbia University
	1150 Amsterdam Avenue, MC 4972
	New York NY 10027

11. Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages, or Other Securities. If none, check box → ☐ None

Full Name	Complete Mailing Address

12. The purpose, function, and nonprofit status of this organization and the exempt status for federal income tax purposes:
☑ Has Not Changed During Preceding 12 Months
☐ Has Changed During Preceding 12 Months (Publisher must submit explanation of change with this statement)

PS Form 3526, October 1999 — See Instructions on Reverse

13. Publication Name: The Journal of Philosophy	14. Issue Date for Circulation Data Below: September 2003	
15. Extent and Nature of Circulation	Average No. Copies Each Issue During Preceding 12 Months	No. Copies of Single Issue Published Nearest to Filing Date
a. Total Number of Copies (Net press run)	3500	3500
b. Paid Circulation (By Mail and Outside the Mail) (1) Mailed Outside-County Paid Subscriptions Stated on PS Form 3541. (Include advertiser's proof copies and exchange copies)	780	646
(2) Mailed In-County Paid Subscriptions Stated on PS Form 3541. (Include advertiser's proof copies and exchange copies)	2490	2239
(3) Paid Distribution Outside the Mails Including Sales Through Dealers and Carriers, Street Vendors, Counter Sales, and Other Paid Distribution Outside USPS®		
(4) Paid Distribution by Other Classes of Mail Through the USPS (e.g. First-Class Mail®)		
c. Total Paid and/or Requested Circulation (Sum of 15b. (1), (2), (3), and (4))	3180	2884
d. Free Distribution by Mail (Samples, complimentary, and other free) (1) Free or Nominal Rate Outside-County Copies included on PS Form 3541		
(2) Free or Nominal Rate In-County Copies included on PS Form 3541		
(3) Free or Nominal Rate Copies Mailed at Other Classes Through the USPS (e.g. First-Class Mail)		
(4) Free or Nominal Rate Distribution Outside the Mail (Carriers or other means)	150	150
e. Total Free or Nominal Rate Distribution (Sum of 15d. (1), (2), (3) and (4))	150	150
f. Total Distribution (Sum of 15c. and 15e.)	3330	3034
g. Copies not Distributed (See Instructions to Publishers #4 (page #3))	170	466
h. Total (Sum of 15f. and g.)	3500	3500
i. Percent Paid and/or Requested Circulation (15c. divided by 15f. times 100)	93.34	95 06

16. Publication of Statement of Ownership
☑ Publication required. Will be printed in the 12/2003 issue of this publication. ☐ Publication not required.

17. Signature and Title of Editor, Publisher, Business Manager, or Owner
John Searle, Managing Editor — Date: 10/14/03

I certify that all information furnished on this form is true and complete. I understand that anyone who furnishes false or misleading information on this form or who omits material or information requested on the form may be subject to criminal sanctions (including fines and imprisonment) and/or civil sanctions (including multiple damages and civil penalties).

Instructions to Publishers

1. Complete and file one copy of this form with your postmaster annually on or before October 1. Keep a copy of the completed form for your records.
2. In cases where the stockholder or security holder is a trustee, include in items 10 and 11 the name of the person or corporation for whom the trustee is acting. Also include the names and addresses of individuals who are stockholders who own or hold 1 percent or more of the total amount of bonds, mortgages, or other securities of the publishing corporation. In item 11, if none, check the box. Use blank sheets if more space is required.
3. Be sure to furnish all circulation information called for in item 15. Free circulation must be shown in items 15d, e, and f.
4. Item 15h., Copies not Distributed, must include (1) newsstand copies originally stated on Form 3541, and returned to the publisher, (2) estimated returns from news agents, and (3), copies for office use, leftovers, spoiled, and all other copies not distributed.
5. If the publication had Periodicals authorization as a general or requester publication, this Statement of Ownership, Management, and Circulation must be published. It must be printed in any issue in October or, if the publication is not published during October, the first issue printed after October.
6. In item 16, indicate the date of the issue in which this Statement of Ownership will be published.
7. Item 17 must be signed.

Failure to file or publish a statement of ownership may lead to suspension of Periodicals authorization.

PS Form 3526, October 1999 (Reverse)

New Perspectives



Rosenzweig and Heidegger
Between Judaism and German Philosophy
PETER ELI GORDON

"Though these two thinkers never met they...shared many of the same views. [Gordon demonstrates] that Rosenzweig should be seen as a late heir to [the post-Hegelian] German philosophical tradition, and not simply as a Jewish thinker."
—*New York Review of Books*

Weimar and New German Cultural Criticism

$65.00
HARDCOVER



Gadamer's Repercussions
Reconsidering Philosophical Hermeneutics
BRUCE KRAJEWSKI, EDITOR

As an investigation and questioning of this profoundly influential philosopher, this collection will become a catalyst for any future rethinking of philosophical hermeneutics, as well as for rereading Gadamer

$65 00 HARDCOVER

NEW IN PAPERBACK

TENTH ANNIVERSARY EDITION
Unbearable Weight
Feminism, Western Culture, and the Body
SUSAN BORDO
New Preface by the Author, New Foreword by Leslie Heywood

"A masterpiece of complex and nuanced thinking not only about a significant problem that faces women but about our culture "
—Susan Griffin, author of *The Book of Courtesans*

$19 95 PAPERBACK

Observations on the Feeling of the Beautiful and Sublime
IMMANUEL KANT
John T. Goldthwait, Translator

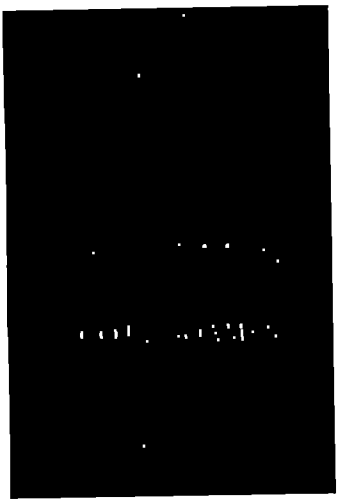
"What emerges from [the] pages is that far from being a dry-as-dust pedant, Kant was a man of warmth, feeling, and humor, who possessed an acute sensitivity for the different shades of aesthetic experience." —*Philosophical Books*

$16.95 PAPERBACK

At bookstores or order (800) 822-6657
www.ucpress.edu

California
UNIVERSITY of CALIFORNIA PRESS

Important



Philosophical Analysis in the Twentieth Century

Scott Soames

This is a major, wide-ranging history of analytic philosophy since 1900, told by one of the tradition's leading contemporary figures. The first volume takes the story from 1900 to mid-century. The second brings the history up to date.

Soames has been at the center of some of the tradition's most important debates and he writes with exceptional ease about its often complex ideas.

"A history of analytic philosophy and an excellent piece of analytic philosophy in its own right. We can all benefit from Soames's discussion of the central issues that have shaped the subject and his assessment of what we have achieved and where we might have gone wrong."—Kit Fine, New York University

Volume 1. *The Dawn of Analysis*
Cloth $35.00 ISBN 0-691-11573-7

Volume 2: *The Age of Meaning*
Cloth $35.00 ISBN 0-691-11574-5

New in paperback

Descartes's Method of Doubt

Janet Broughton

"This stunning work is without question a major contribution to Cartesian studies, to the field of early modern philosophy, and to general epistemology—original, provocative, and philosophically interesting. The writing is everywhere precise and elegant, but straightforward and conversational at the same time. It is a brilliant piece of philosophy."—Edwin McCann, University of Southern California

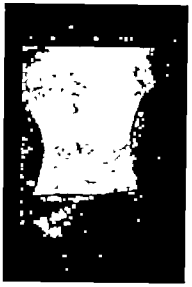
Paper $16.95 ISBN 0-691-11732-2
Due December

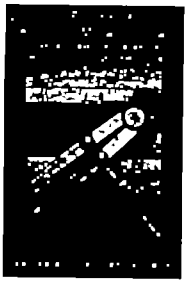


PRINCETON University Press

800-777-4726 • READ EXCERPTS ONLINE
WWW.PUP.PRINCETON.EDU

New from
Oxford







Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556. Visit our website at www.oup.com/us

OXFORD

African Philosophy
New and Traditional Perspectives
Edited by LEE M. BROWN
The primary focus of this collection is on traditional African conceptions of topics like mind, person, personal identity, truth, knowledge, understanding, objectivity, destiny, free will, causation, and reality. The essays in this collection ultimately try to bring a more rigorous conception of African philosophy into fruitful contact with Western philosophical concerns, specifically in the philosophies of psychology, mind, science, and language, as well as in metaphysics and epistemology.
January 2004 paper $19.95 cloth $50.00

Setting the Moral Compass
Essays by Women Philosophers
Edited by CHESHIRE CALHOUN
Setting the Moral Compass brings together the (largely unpublished) work of nineteen women moral philosophers who tackle several wide-ranging projects: they develop an ethics for ordinary life and vulnerable persons; they examine the question of what we ought to do for each other; they highlight the moral significance of inhabiting a shared social world; they reveal the complexities of moral negotiations, and finally they show us the place of emotion in moral life.
December 2003 paper $27.50 cloth $85.00

The Oxford Handbook of Rationality
Edited by ALFRED R. MELE and PIERS RAWLING
Rationality has long been a central topic in philosophy, crossing standard divisions and categories. *The Oxford Handbook of Rationality* is an indispensable reference to the current state of play in this vital and interdisciplinary area of study.
December 2003 $75.00

Identifying the Mind
Selected Papers of U. T. Place
the late U. T. PLACE
Edited by GEORGE GRAHAM and ELIZABETH R. VALENTINE
This collection focuses on Place's philosophy of mind and his contributions to neighboring issues in metaphysics and epistemology. It includes an autobiographical essay as well as a recent paper on the function and neural location of consciousness.
December 2003 $35.00

The Magic Prism
An Essay in the Philosophy of Language
HOWARD WETTSTEIN
In *The Magic Prism*, Wettstein argues that Wittgenstein, a figure with whom the critics of Frege and Russell are typically unsympathetic, laid the foundation for much of what is revolutionary in recent developments in the movement of philosophy of language.
January 2004 $45.00

Announcing

OXFORD

Scholarship Online

Select philosophy titles are now online and fully cross-searchable

Ask your librarian to register for a FREE trial

* Includes abstracts for books as well as individual chapters
* Links from bibliographies and footnotes to online content in related journals
* Search options include: full text, subject, keyword, author, title, and publication date or range
* Ideal for online course packs and reading lists, students will be able to link directly via the library or remote access
* Over 200 new titles added each year

The OSO Philosophy module currently offers 240 titles, including the following leading works:

Annas: *The Morality of Happiness*

Austin: *Philosophical Papers*

Bennett: *Learning from Six Philosophers 1* and *2*

Davidson: *Subjective, Intersubjective, Objective*

Dummett: *The Seas of Language*

Field: *Truth and the Absence of Fact*

Fodor: *Concepts*

Foot: *Natural Goodness*

Gauthier: *Morals by Agreement*

Grice: *Aspects of Reason*

Jackson: *From Metaphysics to Ethics*

Mackie: *The Cement of the Universe*

Maddy: *Naturalism in Mathematics*

Nagel: *Equality and Partiality*

Plantinga: *The Nature of Necessity*

Rae: *The Morality of Freedom*

Strawson: *Entity and Identity*

Swinburne: *The Existence of God*

van Fraassen: *The Scientific Image*

Williamson: *Knowledge and its Limits*

Oxford Scholarship Online is available by annual subscription either as a complete collection or by individual discipline. Free trials are available to institutions. Please ask your librarian to register now.

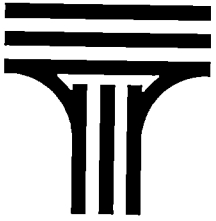
Institutions in North and South America:
By email: onlinesubscriptions@oup-usa.org
By fax: 212 726 6476
By phone: 800 334 4249 ext 6484

Institutions outside North and South America:
By email: onlineproducts@oup.com
By fax: +44 (0) 1865 353308

For further information—including a complete list of available titles, as well as an online demonstration—please visit **www.oxfordscholarship.com**

Classics in Philosophy

NEW FROM THOEMMES PRESS
www.thoemmes.com



LEVIATHAN
Thomas Hobbes
Edited by Karl Schuhmann and G. A. J. Rogers
800 PAGES, 2 VOLUMES
CLOTH $225.00

AN ESSAY CONCERNING HUMANE UNDERSTANDING, 1ST AND 5TH EDITIONS
1890 and 1906 Editions
John Locke
980 PAGES, 2 VOLUMES
CLOTH $375.00

LOCKE IN GERMANY
Early German Translations of John Locke, 1709–61
Edited by Konstantin Pollok
2280 PAGES, 8 VOLUMES
CLOTH $1095.00

ESSAYS IN PHILOSOPHY AND SOCIAL POLICY, 1883–1922
Bernard Bosanquet
Edited by William Sweet
1070 PAGES, 3 VOLUMES
CLOTH $375.00

PHILOSOPHICAL WORKS OF DAMARIS, LADY MASHAM
Edited by James G. Buickerood
400 PAGES
CLOTH $125.00

EARLY RESPONSES TO HUME'S LIFE AND REPUTATION
Edited by James Fieser
800 PAGES, 2 VOLUMES
CLOTH $280.00

THE WORKS OF JAMES HARRIS
Edited by James Harris, Earl of Malmesbury
1168 PAGES, 2 VOLUMES
CLOTH $375.00

WOMEN IN THE ST. LOUIS IDEALIST MOVEMENT, 1860–1925
Dorothy G. Rogers
1169 PAGES, 4 VOLUMES
CLOTH $445.00

DISTRIBUTED BY THE UNIVERSITY OF CHICAGO PRESS
1427 East 60th Street, Chicago, IL 60637 • www.press.uchicago.edu

Knowledge of Things Human and Divine



Vico's New Science and Finnegans Wake

Donald Phillip Verene

"Verene has joined a masterful account of the progress of Vico's thought and writing to a parallel, or overlapping, commentary on the complex and often cryptically concealed Vichian aspects of James Joyce's work. **This book is of the first importance.**"
—Donald R. Kelley $45.00

Spinoza's Book of Life

Freedom and Redemption in the Ethics

Steven B. Smith

"**This important book** dispels a number of major misunderstandings about Spinoza and shows his importance for an understanding of modernity"—David Novak $35.00

The Tragedy of Finitude

Dilthey's Hermeneutics of Life

Jos de Mul

Translated by Anthony Burrett

"**A thorough, insightful account of Dilthey's philosophy**, this book offers many valuable new contributions. De Mul argues effectively for Dilthey's relevance today"
—Rudolf Makkreel

Yale Studies in Hermeneutics $45.00

New in Paper

The Heavenly City of the Eighteenth-Century Philosophers

Second Edition

Carl L. Becker

With a new foreword by Johnson Kent Wright

"**An enduring landmark** in the historiography of the Enlightenment. For three generations its eloquence of style and elegance of concept have provoked learning and creative criticism."—Frank M. Turner

$16.00 A Yale Nota Bene paperback

New in the Rethinking the Western Tradition Series

Groundwork for the Metaphysics of Morals

Immanuel Kant

Edited and translated by Allen W. Wood

With essays by J.B. Schneewind, Marcia Baron, Shelly Kagan, and Allen W. Wood

$35.00 cloth; $12.95 paperback

Reflections on the Revolution in France

Edmund Burke

Edited by Frank M. Turner

With essays by Conor Cruise O'Brien, Darrin M. McMahon, Jack N. Rakove, and Alan Wolfe

$20.00 cloth, $9.95 paperback

Two Treatises of Government *and* A Letter Concerning Toleration

John Locke

Edited by Ian Shapiro

With essays by John Dunn, Ruth Grant, and Ian Shapiro

$30.00 cloth, $14.00 paperback

YALE University Press
yalebooks.com

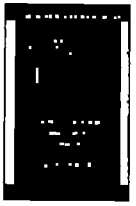
CAMBRIDGE

outstanding scholarship



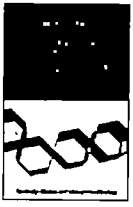
Art versus NonArt
Art out of Mind
Tsion Avital
Contemporary Artists and their Critics
$85.00 Hardback 0-521-82465-6 464pp

John Stuart Mill
A Biography
Nicholas Capaldi
$40.00* Hardback 0-521-62024-4 350pp



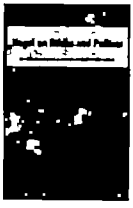
The Mind and Its Stories
Narrative Universals and Human Emotion
Patrick Colm Hogan
Studies in Emotion and Social Interaction
$65.00 Hardback 0-521-82527-X 320pp

The Cambridge Companion to Quine
Edited by Roger F. Gibson, Jr.
Cambridge Companions to Philosophy
$70.00* Hardback 0-521-63056-8 350pp
$25.00* Paperback 0-521-63949-2



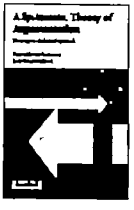
Natural Kinds and Conceptual Change
Joseph LaPorte
Cambridge Studies in Philosophy and Biology
$70.00 Hardback 0-521-82599-7 237pp

A Physicalist Manifesto
Thoroughly Modern Materialism
Andrew Melnyk
Cambridge Studies in Philosophy
$70.00 Hardback 0-521-82711-6 340pp



Hegel on Ethics and Politics
Edited by Robert B. Pippin and Otfried Hoffe
The German Philosophical Tradition
$65.00 Hardback 0-521-81814-1 360pp

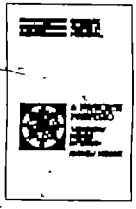
The Stag Hunt and the Evolution of Social Structure
Brian Skyrms
$55.00 Hardback 0-521-82651-9 168pp
$20.00 Paperback 0-521-53392-9

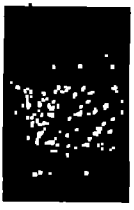


A Systematic Theory of Argumentation
The Pragma-dialectical Approach
Frans H. van Eemeren and Rob Grootendorst
$58.00 Hardback 0-521-83075-3 224pp
$22.00 Paperback 0-521-53772-X









*Prices subject to change

www.cambridge.org

CAMBRIDGE UNIVERSITY PRESS